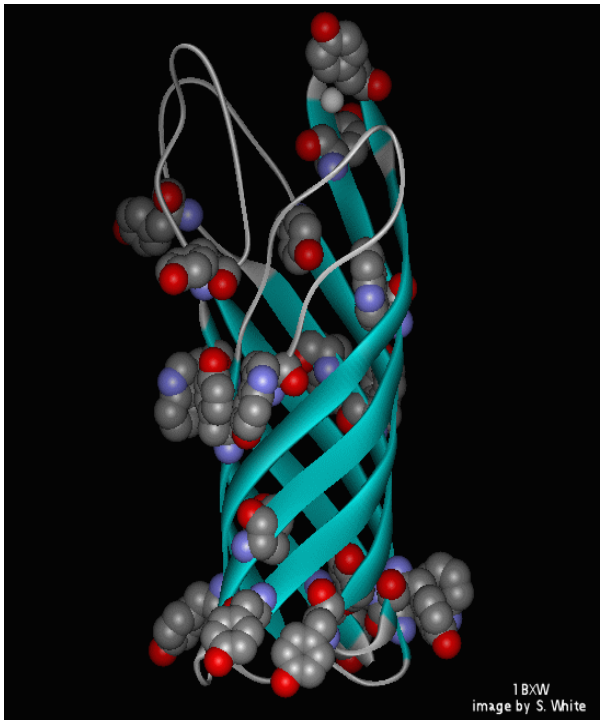
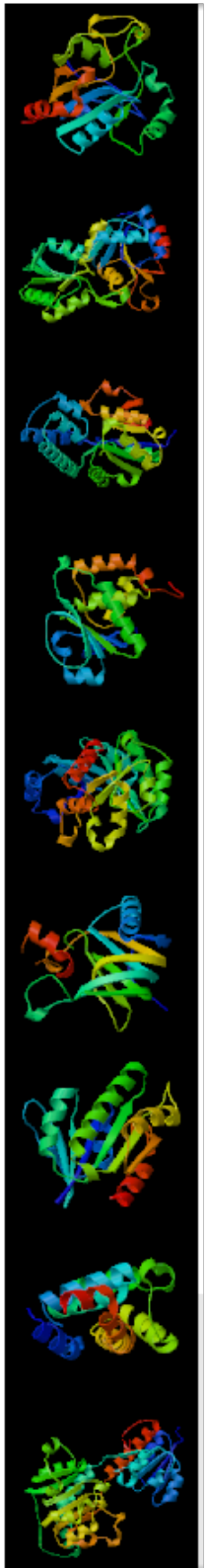
Protein Structure Prediction

- Why ?
- Introduction to protein structure predictions
- Secondary structure prediction (Mount: 455-468)
  - Chou-Fasman
  - GOR-III method
  - PhD method
  - Nearest Neighbor methods
  - State of the art methods
- Molecular modelling - intro



Arne Elofsson (arne@bioinfo,se)



Why do we need structure prediction?

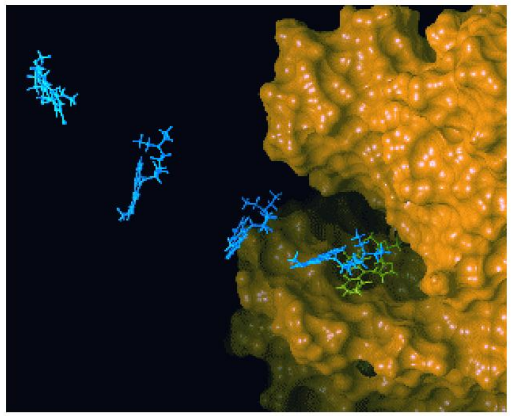
- 3D structure give clues to function:
  - active sites, binding sites, conformational changes...
  - structure and function conserved more than sequence
  - 3D structure determination is difficult, slow and expensive
  - Intellectual challenge, Nobel prizes etc...
  - Engineering new proteins

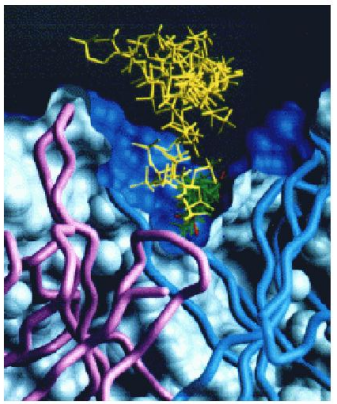
The Use of Structure

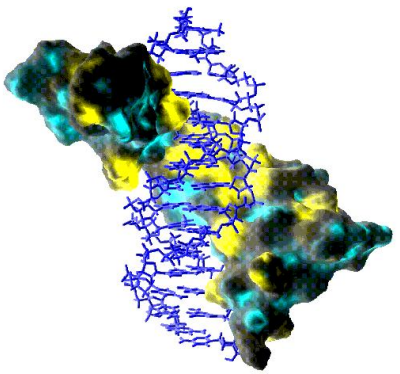
Major Application I: Designing Drugs

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).

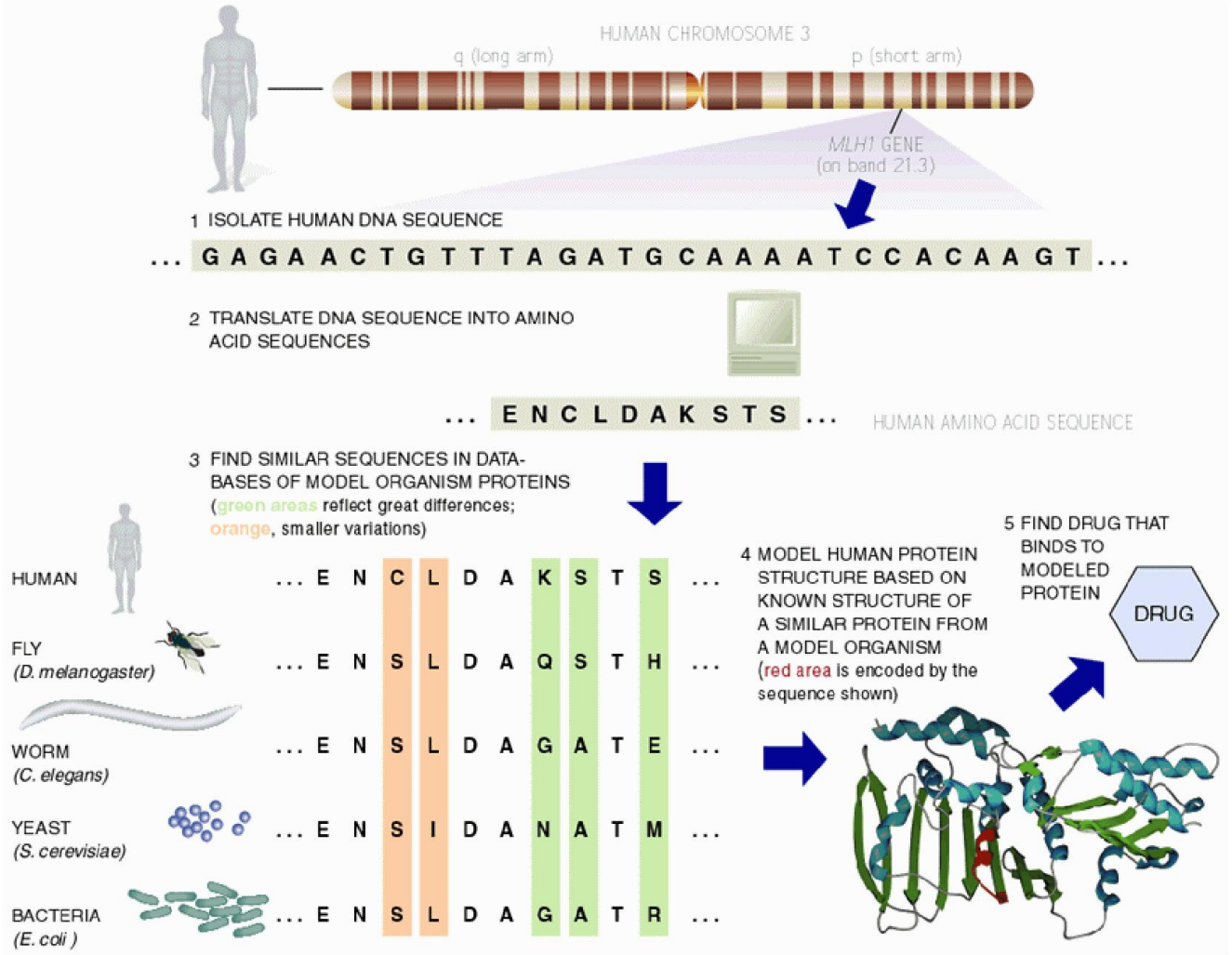






The Use of Structure

Major Application II: Finding Homologs

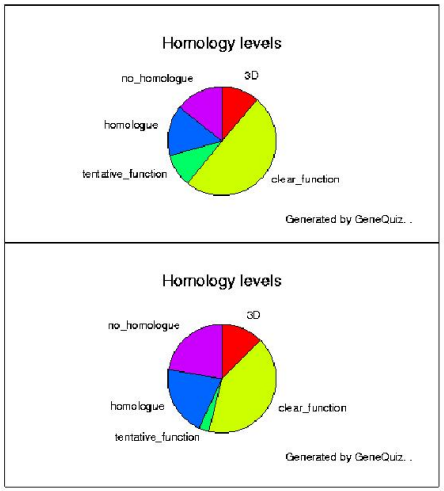


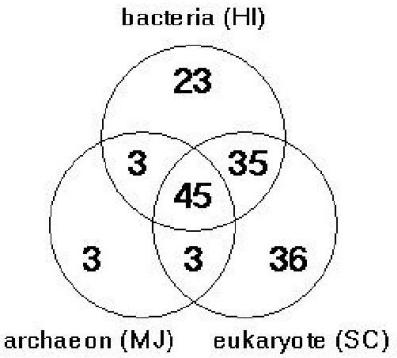
The Use of Structure

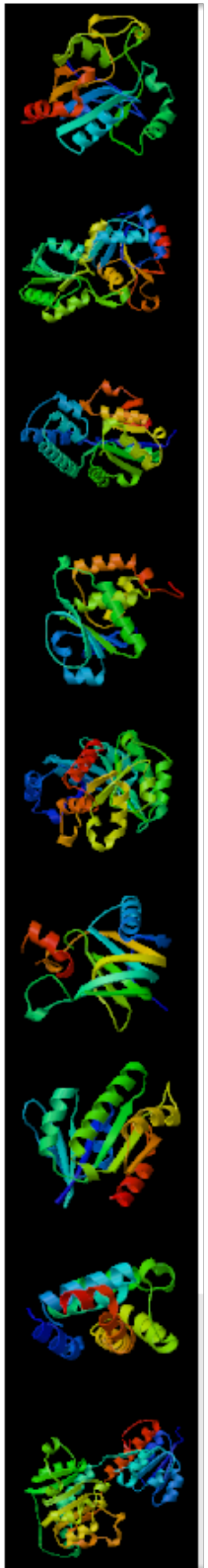
Major Application I|I:
Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics

(Clock figures, yeast v. Synechocystis,
adapted from GeneQuiz Web Page, Sander Group, EBI)



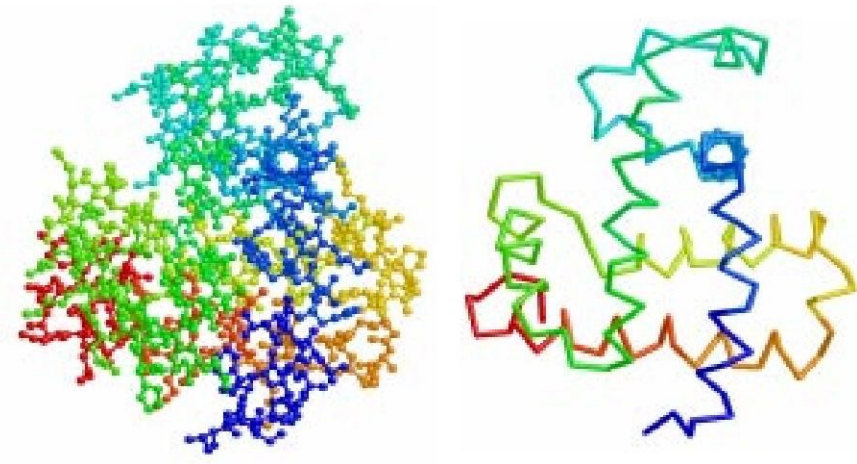


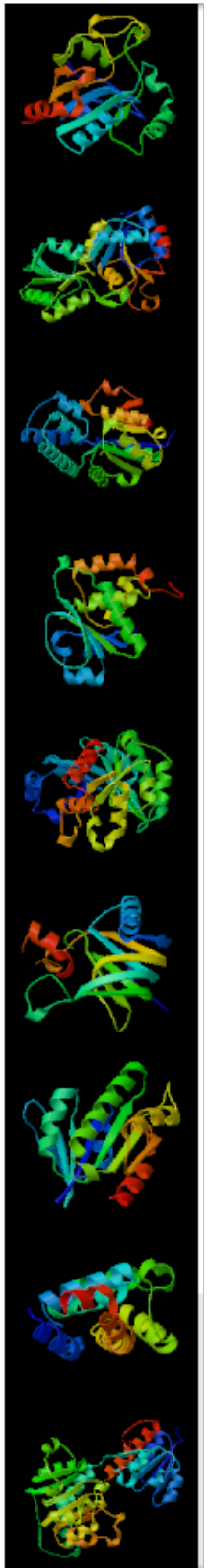


It's not that simple...

- Amino acid sequence contains all the information for 3D structure (experiments of Anfinsen, 1970's)
- But, there are thousands of atoms, rotatable bonds, solvent and other molecules to deal with...
- Levinthal's paradox

Sperm Whale Myoglobin





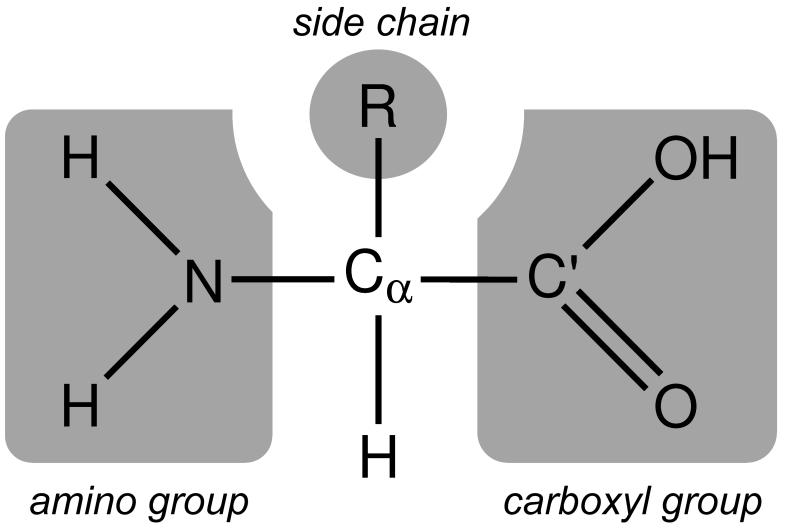
Remember..

- All which we study is an abstraction to make comprehension of a complex entity more straightforward
- We think of structures as static entities, but they are dynamic, sometimes to the point of being ill-definable – function requires this flexibility
- The more we have the more we know

Arne Elofsson (arne@bioinfo,se)

Primary Structure - Amino Acids

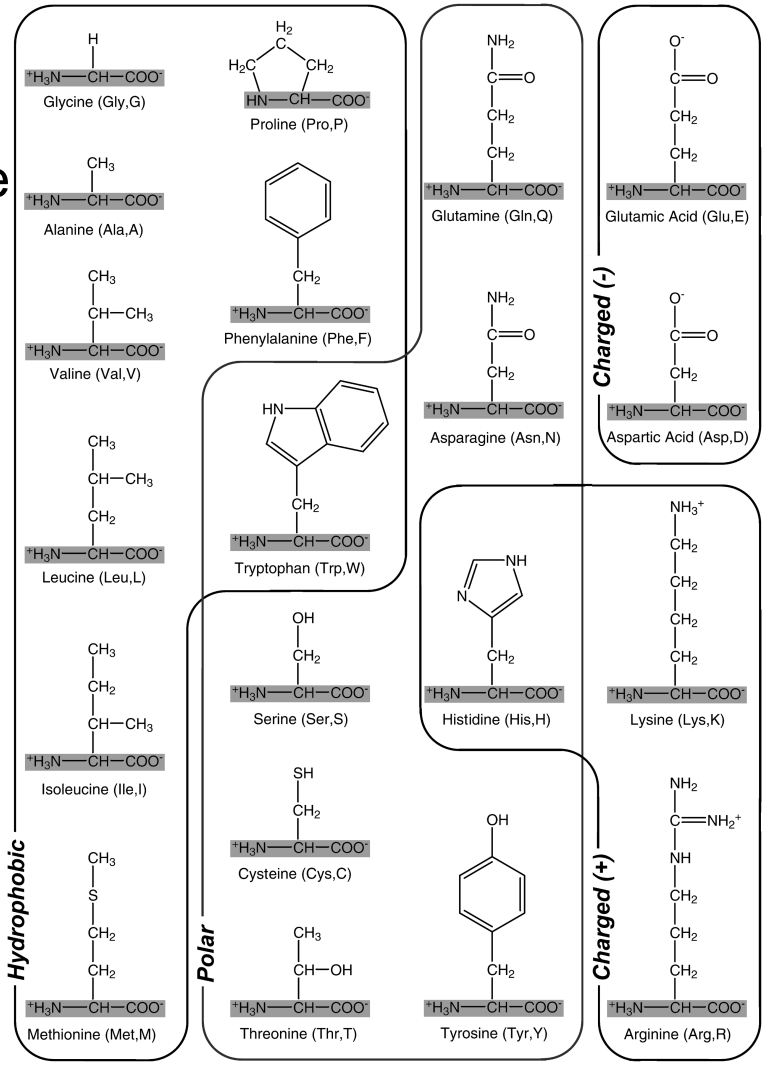
- It is the amino acid sequence that "exclusively" determines the 3D structure of a protein
- 20 amino acids – modifications do occur post protein synthesis



Arne Elofsson (arne@bioinfo,se)

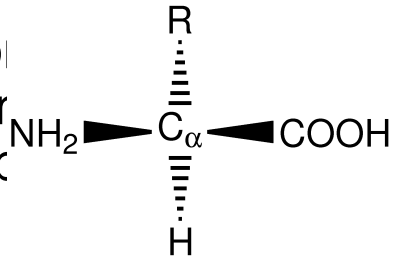
Amino Acids Continued…

- It is the properties of the R group that determine the property of the aa and ultimately the protein
- Different schemes exist for describing the properties Willie Taylor's scheme is often employed in bioinformatics analyses
- Hydrophobicity, polarity and charge are common measures
- Learn the amino acid codes, structures and properties!

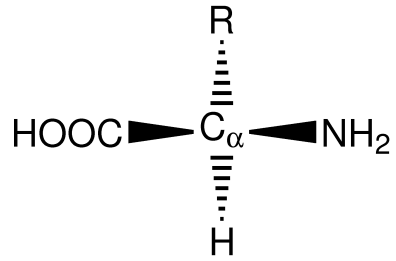


Amino Acids Continued…

- Chirality – amino acids are enatiomorphs, that is mirror images exist – only the L(S) form is found in naturally forming proteins. Some enzymes can produce D(R) amino acids
- Think about a data structure fo this information – annotation ar a validation procedure should b included
- Think about systematic versus common nomenclature



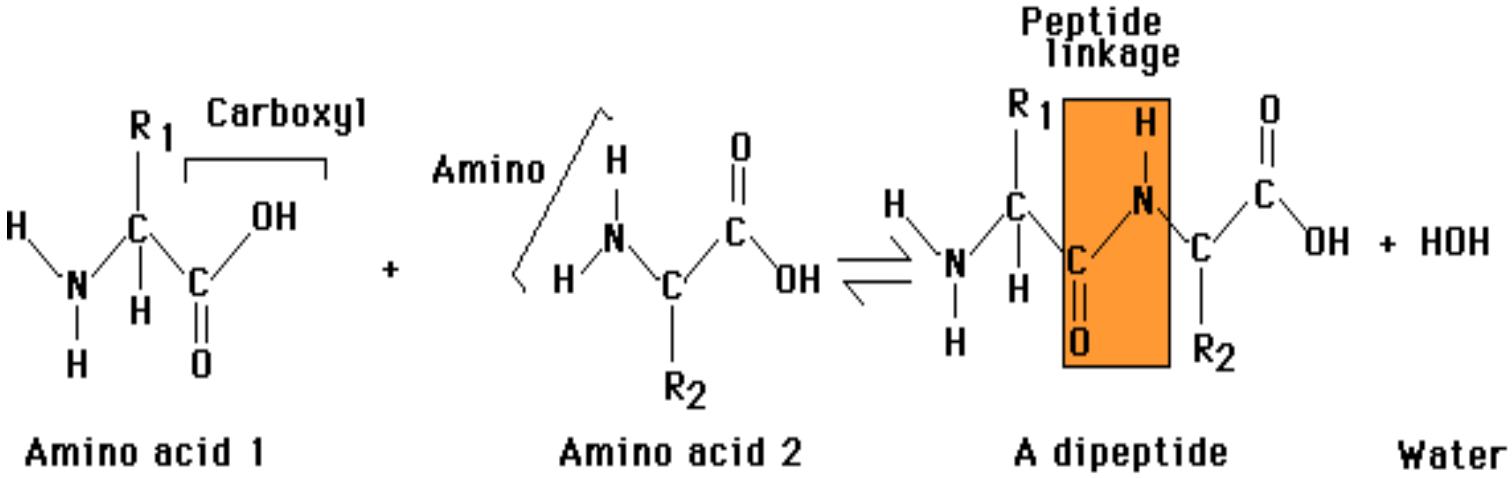
L-amino acid



D-amino acid

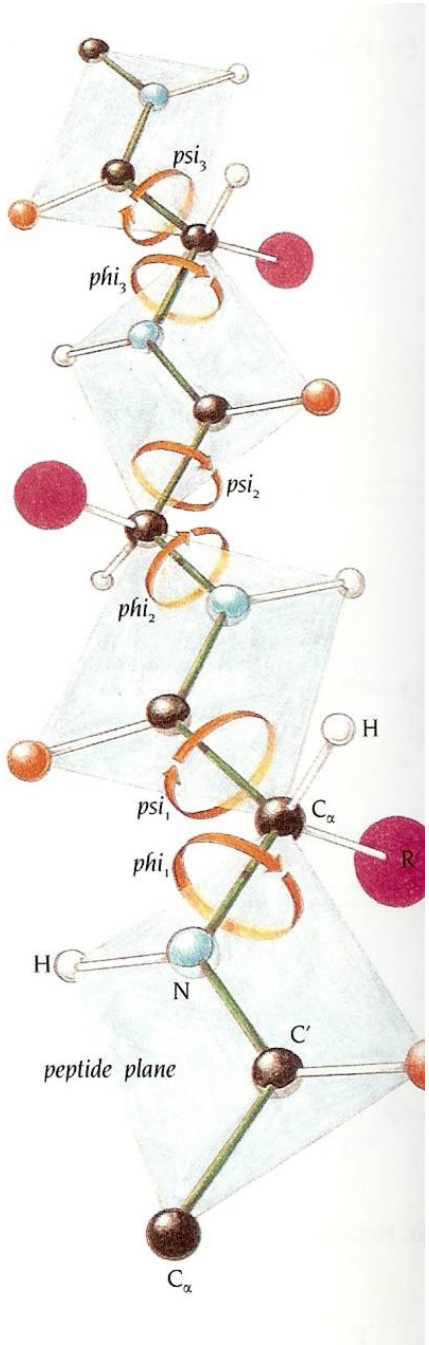
Peptide Bond Formation

- Individual amino acids form a polypeptide chain
- Such a chain is a component of a hierarchy for describing macromolecular structure
- The chain has its own set of attributes
- The peptide linkage is planar and rigid



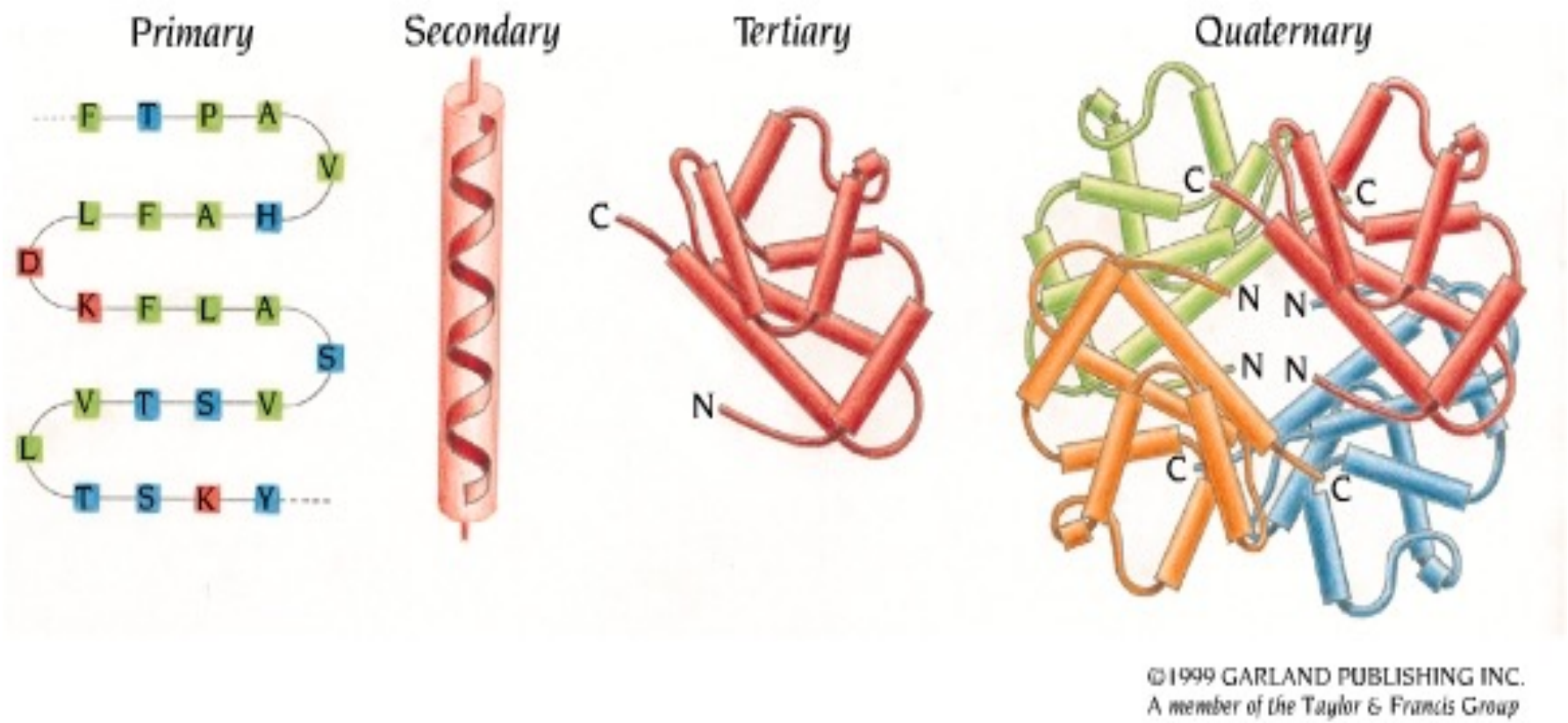
Geometry of the Chain

- A dihedral angle is the angle between two planes defined by 4 atoms – 123 make one plane 234 the other
- Omega is the rotation around the peptide bond C_n – N_{n+1} – it is planar and is 180 under ideal conditions
- Phi is the angle around N – Calpha
- Psi is the angle around Calpha C'
- The values of phi and psi are constrained to certain values based on steric clashes of the R group. Thus these values show characteristic patterns as defined by the Ramachandran plot

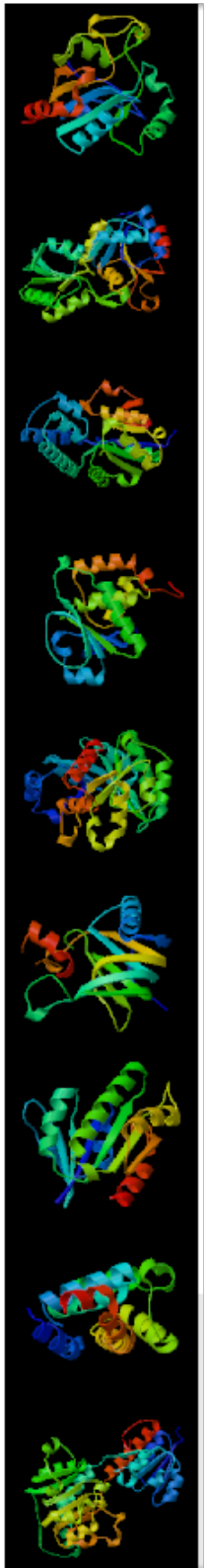


From Brandon and Tooze

4 Basic Levels of Protein Structure

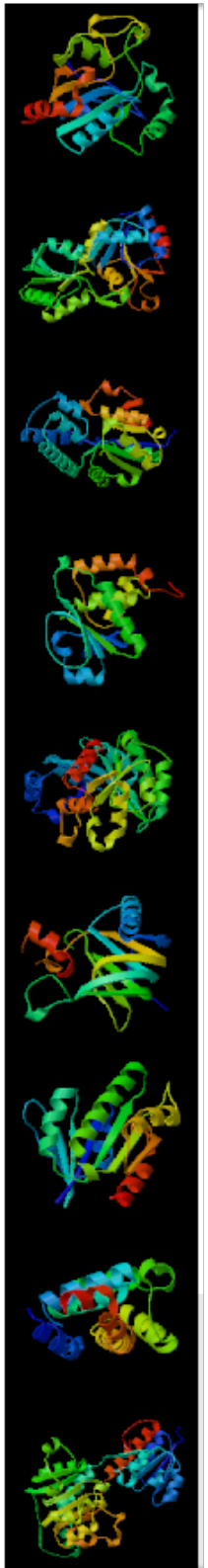


Arne Elofsson (arne@bioinfo,se)



Summary of the four main approaches to structure prediction. Note that there are overlaps between nearly all categories.

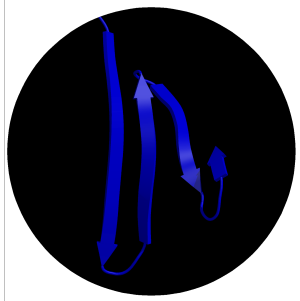
Method	Knowledge	Approach	Difficulty	Usefulness
Comparative modelling (Homology modelling)	Proteins of known structure	Identify related structure with sequence methods, copy 3D coords and modify where necessary	Relatively easy	Very, if sequence identity drug design
Fold recognition	Proteins of known structure	Same as above, but use more sophisticated methods to find related structure	Medium	Limited due to poor models
Secondary structure prediction	Sequence-structure statistics	Forget 3D arrangement and predict where the helices/strands are	Medium	Can improve alignments, fold recognition, *ab initio*
ab initio tertiary structure prediction	Energy functions, statistics	Simulate folding, or generate lots of structures and try to pick the correct one	Very hard	X



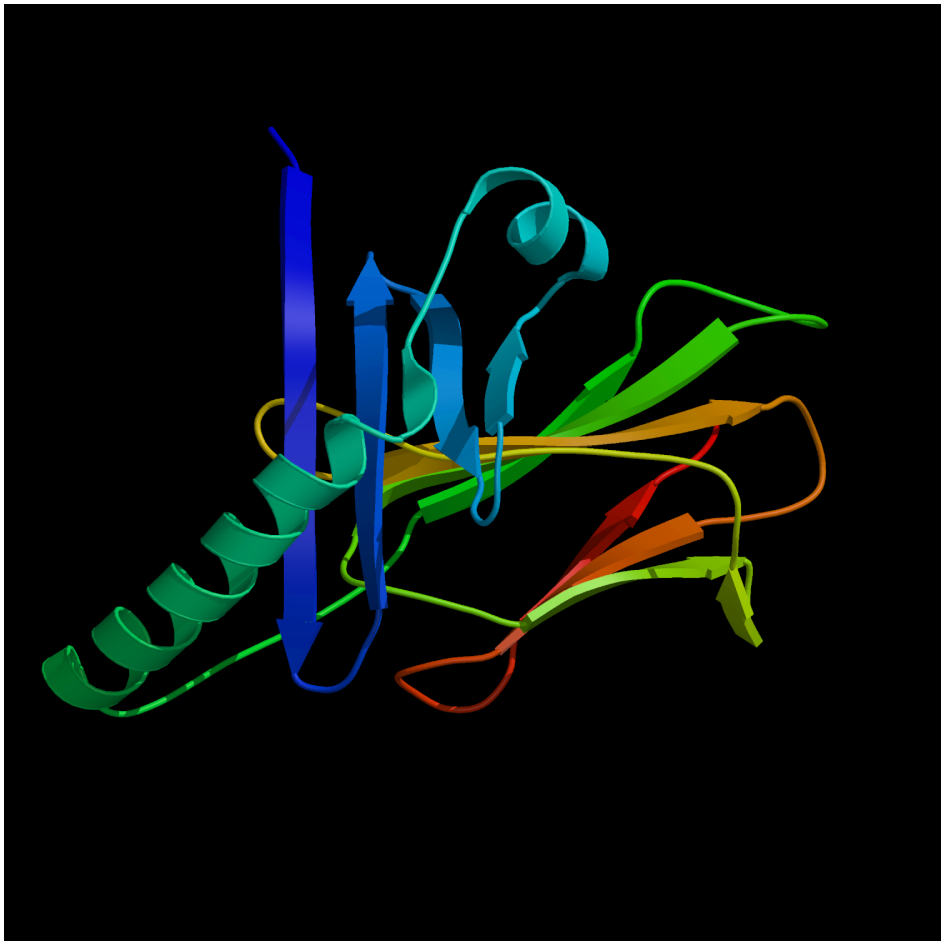
Secondary structure predictions

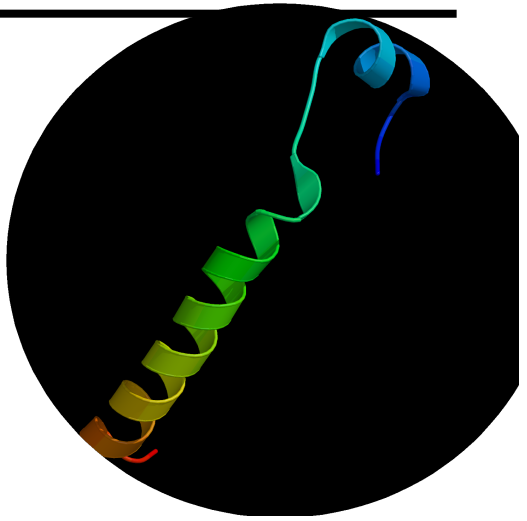
- Ignore 3D, it's too hard!
  - *Usually* concentrate on helix, strand and ``coil".
- Pattern recognition, but which patterns?
- some amino acids have preferences for helix or strand; due to geometry and hydrogen bonding
- spatial (along sequence) patterns, alternating hydrophobics (helical wheel)
- conservation (down alignment) in different members of protein family; insertions and deletions
- Three main generations/stages in SSP method development since 1970's.

Secondary Structure Elements



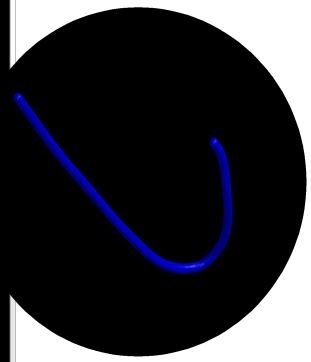
ß-strand



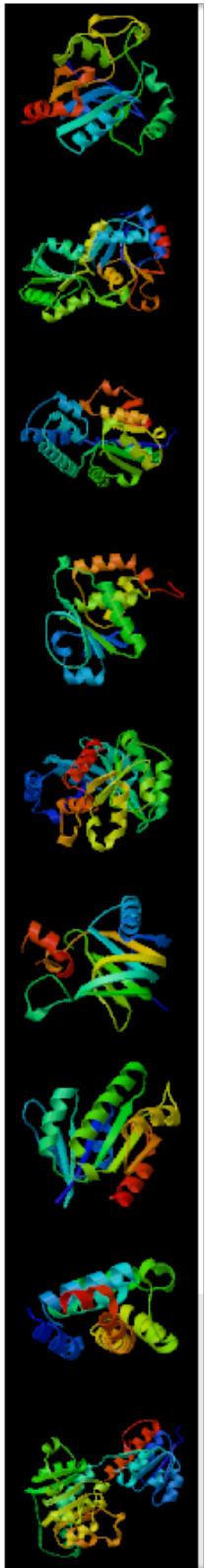


Helix

Turn

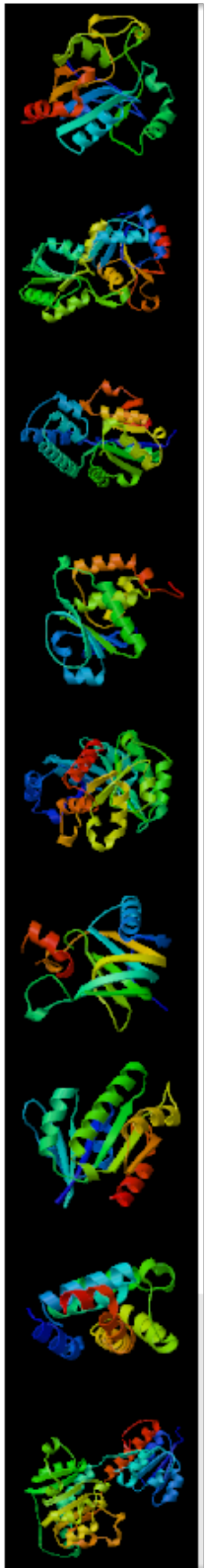


Bend

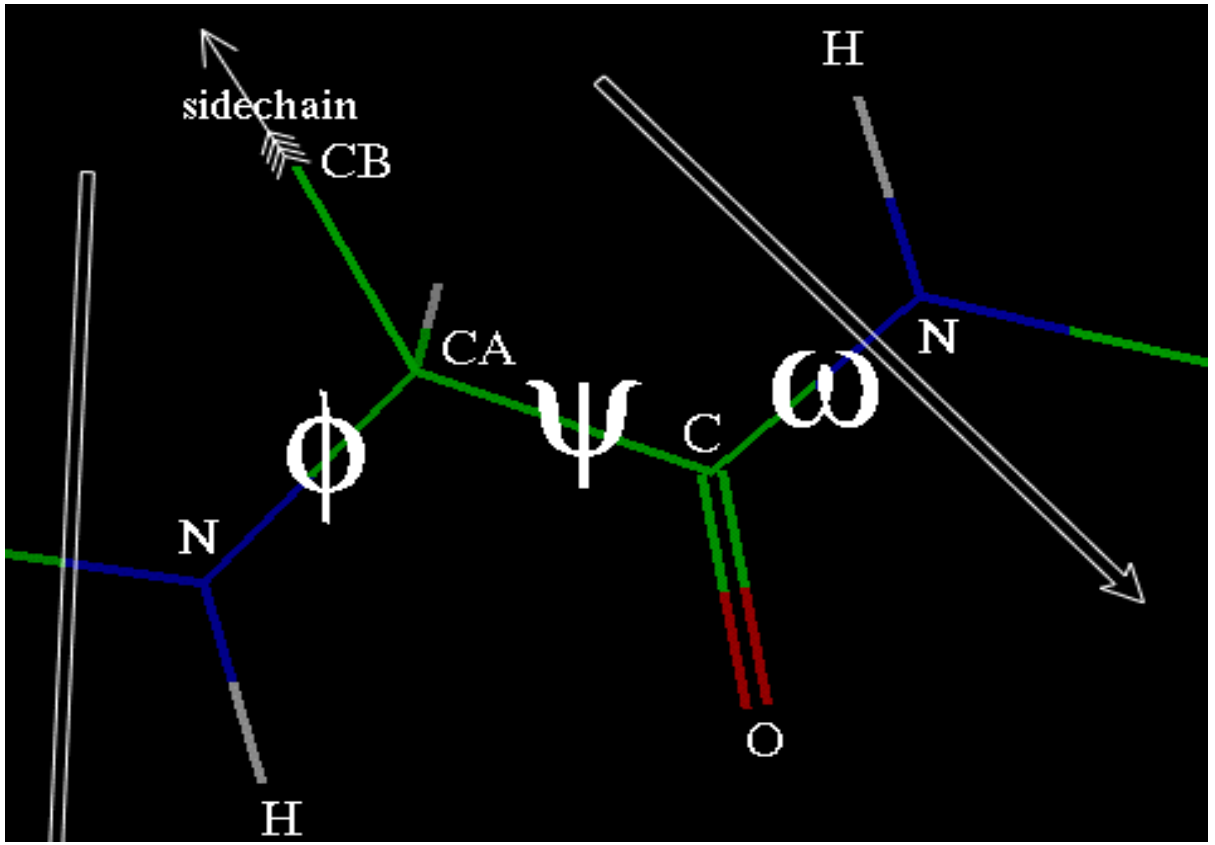


What is ``known secondary structure"?

- Of critical importance in training/assessment of SSP methods
- Can be defined:
- visually by structural biologist
- by geometric and chemical criteria (, angles, distances between atoms, hydrogen bonds...) by programs like DSSP and STRIDE



Dihedral Angles

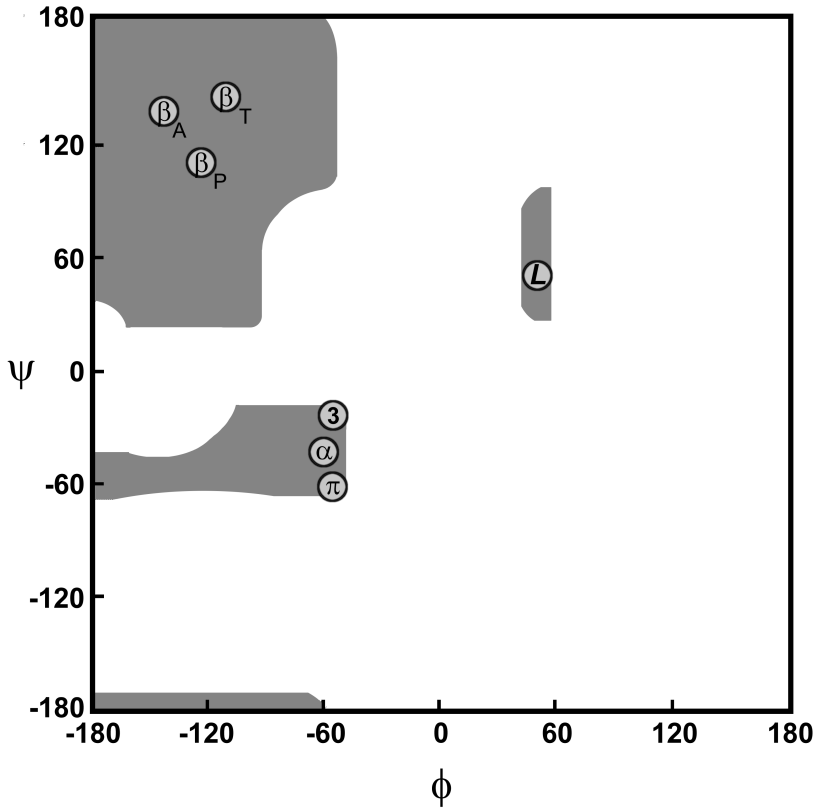


From http://www.imb-jena.de

phi - dihedral angle about the N-C*alpha* bond
psi - dihedral angle about the C*alpha*-C bond
omega - dihedral angle about the C-N (peptide) bond

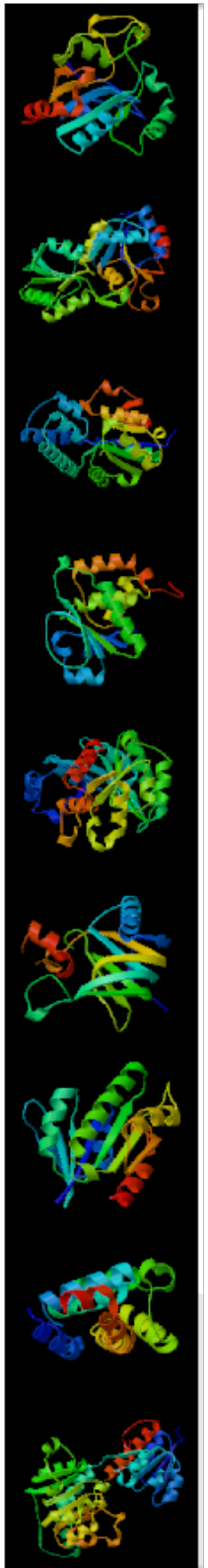
Ramachandran Plot

- Shows allowed and disallowed regions
- Gly and Pro are acceptions: Gly has nc limitation; Pro is constrained by the fac its side chain binds back to the main chain



Gray = allowed conformations. β_A, antiparallel b sheet; β_P, parallel b sheet; β_T, twisted b sheet (parallel or anti-parallel); α, right-handed α helix; ***L***, left-handed helix; **3**, 3_{10} helix; **p,** Π helix.

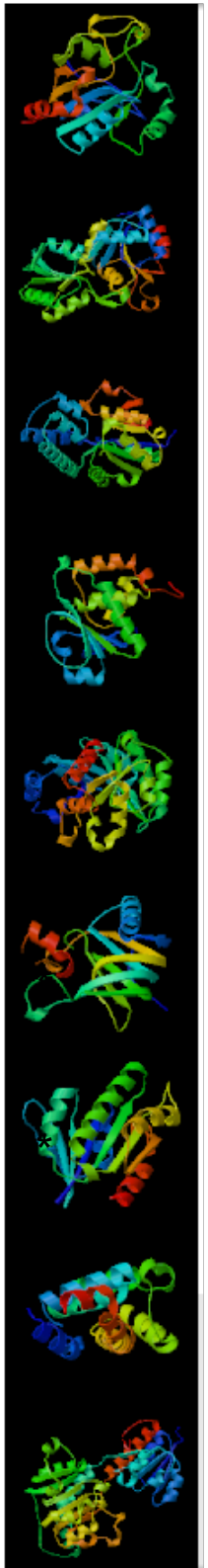
Arne Elofsson (arne@bioinfo,se)



Automatic assignment programs

- DSSP (http://www.cmbi.kun.nl/gv/dssp/)
- STRIDE (http://www.hgmp.mrc.ac.uk/Registered/Option/stride.html)

```
#  RESIDUE AA STRUCTURE BP1 BP2  ACC    N-H-->O    O-->H-N    N-H-->O    O-->H-N     TCO  KAPPA ALPHA   PHI    PSI    X-CA   Y-CA   Z-CA
  1    4 A E             0   0  205    0, 0.0     2,-0.3     0, 0.0     0, 0.0    0.000 360.0 360.0 360.0 113.5    5.7   42.2   25.1
  2    5 A H         -   0   0  127    2, 0.0     2,-0.4    21, 0.0    21, 0.0   -0.987 360.0-152.8-149.1 154.0    9.4   41.3   24.7
  3    6 A V         -   0   0   66   -2,-0.3    21,-2.6     2, 0.0     2,-0.5   -0.995   4.6-170.2-134.3 126.3   11.5   38.4   23.5
  4    7 A I   E     -A  23  0A 106   -2,-0.4     2,-0.4    19,-0.2    19,-0.2   -0.976  13.9-170.8-114.8 126.6   15.0   37.6   24.5
  5    8 A I   E     -A  22  0A  74   17,-2.8    17,-2.8    -2,-0.5     2,-0.9   -0.972  20.8-158.4-125.4 129.1   16.6   34.9   22.4
  6    9 A Q   E     -A  21  0A  86   -2,-0.4     2,-0.4    15,-0.2    15,-0.2   -0.910  29.5-170.4 -98.9 106.4   19.9   33.0   23.0
  7   10 A A   E     +A  20  0A  18   13,-2.5    13,-2.5    -2,-0.9     2,-0.3   -0.852  11.5 172.8-108.1 141.7   20.7   31.8   19.5
  8   11 A E   E     +A  19  0A  63   -2,-0.4     2,-0.3    11,-0.2    11,-0.2   -0.933   4.4 175.4-139.1 156.9   23.4   29.4   18.4
  9   12 A F   E     -A  18  0A  31    9,-1.5     9,-1.8    -2,-0.3     2,-0.4   -0.967  13.3-160.9-160.6 151.3   24.4   27.6   15.3
 10   13 A Y   E     -A  17  0A  36   -2,-0.3     2,-0.4     7,-0.2     7,-0.2   -0.994  16.5-156.0-136.8 132.1   27.2   25.3   14.1
 11   14 A L   E  >> -A  16  0A  24    5,-3.2     4,-1.7    -2,-0.4     5,-1.3   -0.929  11.7-122.6-120.0 133.5   28.0   24.8   10.4
 12   15 A N   T  45S+    0   0   54   -2,-0.4    -2, 0.0     2,-0.2     0, 0.0   -0.884  84.3   9.0-113.8 150.9   29.7   22.0    8.6
 13   16 A P   T  45S+    0   0  114    0, 0.0    -1,-0.2     0, 0.0    -2, 0.0   -0.963 125.4  60.5 -86.5   8.5   32.0   21.6    6.8
 14   17 A D   T  45S-    0   0   66    2,-0.1    -2,-0.2     1,-0.1     3,-0.1    0.752  89.3-146.2 -64.6 -23.0   33.0   25.2    7.6
 15   18 A Q   T  <5 +    0   0  132   -4,-1.7     2,-0.3     1,-0.2    -3,-0.2    0.936  51.1 134.1  52.9  50.0   33.3   24.2   11.2
 16   19 A S   E    < +A  11  0A  44   -5,-1.3    -5,-3.2     2, 0.0     2,-0.3   -0.877  28.9 174.9-124.8 156.8   32.1   27.7   12.3
 17   20 A G   E     -A  10  0A  28   -2,-0.3     2,-0.3    -7,-0.2    -7,-0.2   -0.893  15.9-146.5-151.0-178.9   29.6   28.7   14.8
 18   21 A E   E     -A   9  0A  14   -9,-1.8    -9,-1.5    -2,-0.3     2,-0.4   -0.979   5.0-169.6-158.6 146.0   28.0   31.5   16.7
 19   22 A F   E     +A   8  0A   3   12,-0.4    12,-2.3    -2,-0.3     2,-0.3   -0.982  27.8 149.2-139.1 120.3   26.5   32.2   20.1
 20   23 A M   E     -AB  7  30A  0  -13,-2.5   -13,-2.5    -2,-0.4     2,-0.4   -0.983  39.7-127.8-152.1 161.6   24.5   35.4   20.6
 21   24 A F   E     -AB  6  29A 45    8,-2.4     7,-2.9    -2,-0.3     8,-1.0   -0.934  23.9-164.1-112.5 137.7   21.7   37.0   22.6
 22   25 A D   E     -AB  5  27A  6  -17,-2.8   -17,-2.8    -2,-0.4     2,-0.5   -0.948   6.9-165.0-123.7 138.3   18.9   38.9   20.8
 23   26 A F   E >   S-AB  4  26A  76    3,-3.5     3,-2.1    -2,-0.4   -19,-0.2   -0.947  78.4 -27.2-127.3 111.5   16.4   41.3   22.3
 24   27 A D   T 3   S-    0   0   74  -21,-2.6   -20,-0.1    -2,-0.5    -1,-0.1    0.904 128.9 -46.6  50.4  45.0   13.4   42.1   20.2
 25   28 A G   T 3   S+    0   0   20  -22,-0.3     2,-0.4     1,-0.2    -1,-0.3    0.291 118.8 109.3  84.7 -11.1   15.4   41.4   17.0
 26   29 A D   E <   S-B  23  0A 114   -3,-2.1    -3,-3.5   109, 0.0     2,-0.3   -0.822  71.8-114.7-103.1 140.3   18.4   43.4   18.1
 27   30 A E   E     -B   22  0A   8   -2,-0.4    -5,-0.3    -5,-0.2     3,-0.1   -0.525  24.9-177.7 -74.1 127.5   21.8   41.8   19.1
```



Secondary Structure Prediction

- What to predict?
  - All 8 types or pool types into groups

DSSP

Q3

H = alpha helix
G = 3_{10} -helix
I = 5 helix (pi helix)

E = extended strand
B = beta-bridge

T = hydrogen bonded turn
S = bend
C = coil

H

E

C

Secondary Structure Prediction

Straight HEC

- What to predict?
  - All 8 types or pool types into groups

Q3

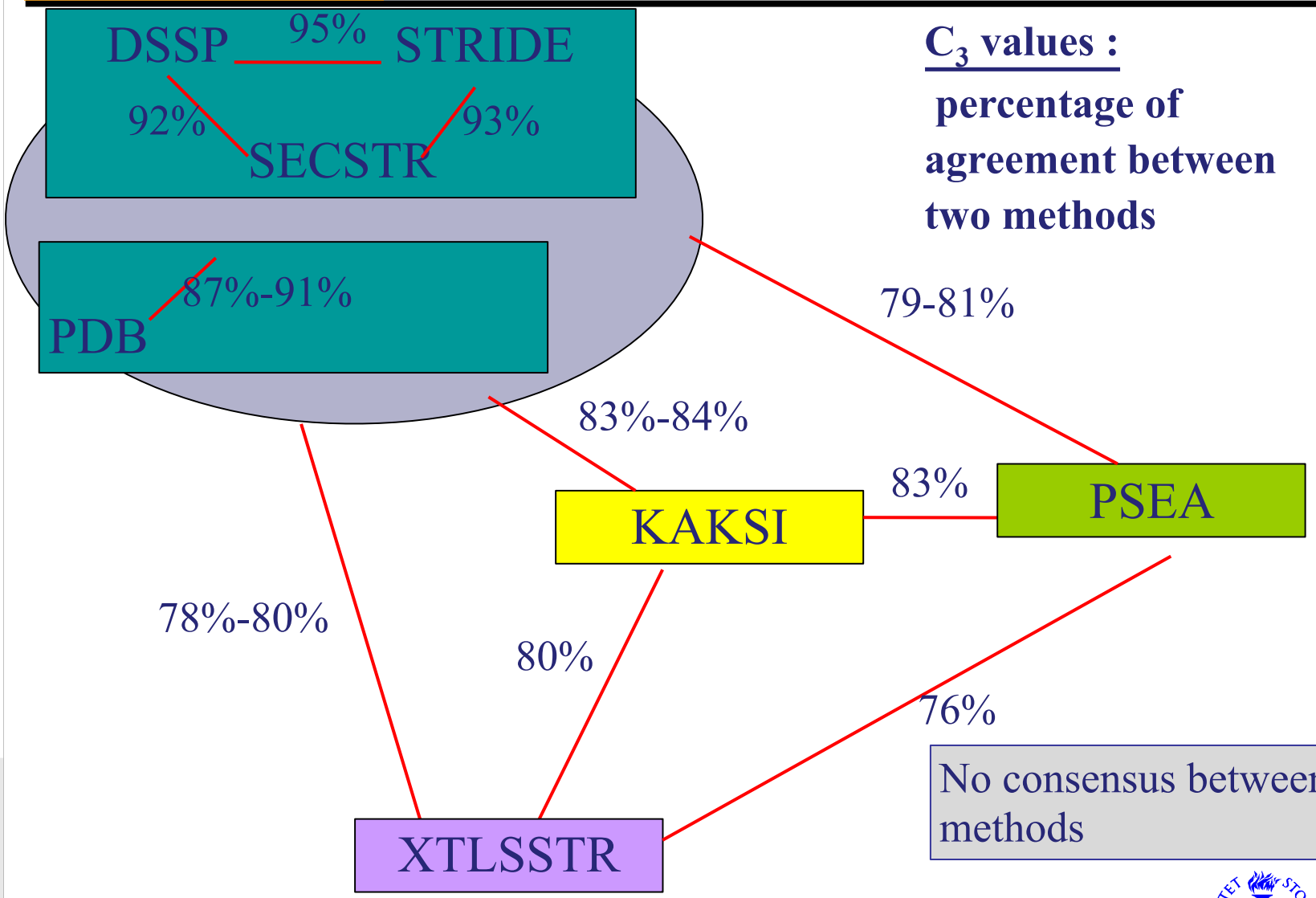
H = alpha helix → H

E = extended strand → E

T = hydrogen bonded turn
S = bend
C = coil
G = 3_{10}-helix
I = 5 helix (pi helix)
B = beta-bridge → C

Secondary structures agreement between programs

Hydrogen bonds



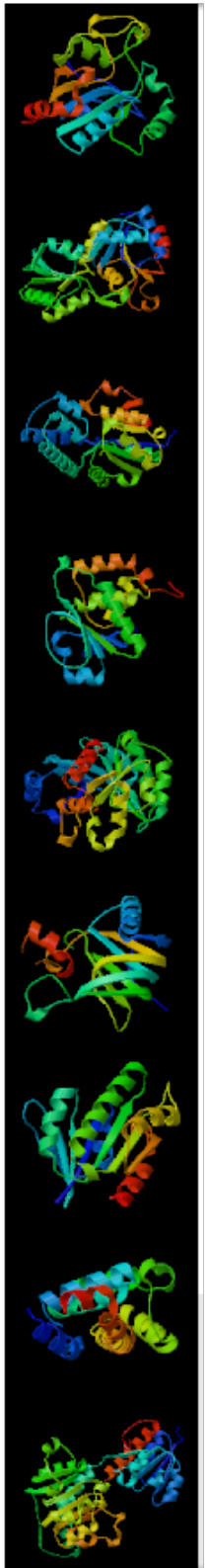
C_3 values :

percentage of agreement between two methods

No consensus between methods

Arne Elofsson (arne@bioinfo.se)

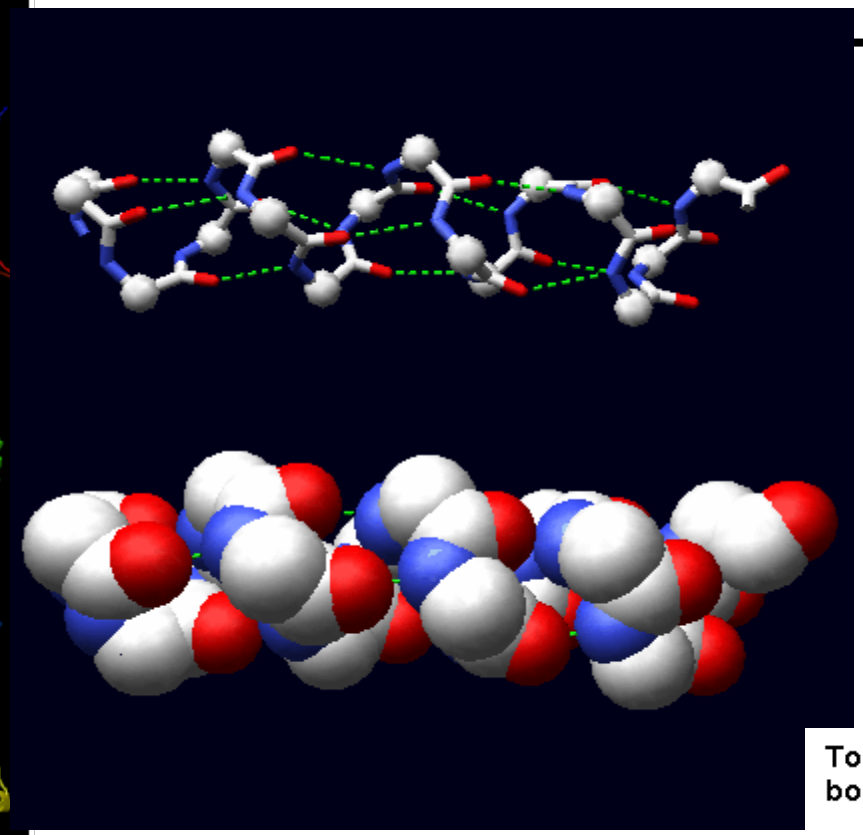
J. Martin, G. Letellier,
J.F. Taly, A. Martin, A.G. de Brevern & J.F. Gibrat, (2005) *BMC Structural Biology*, **5**, 17.

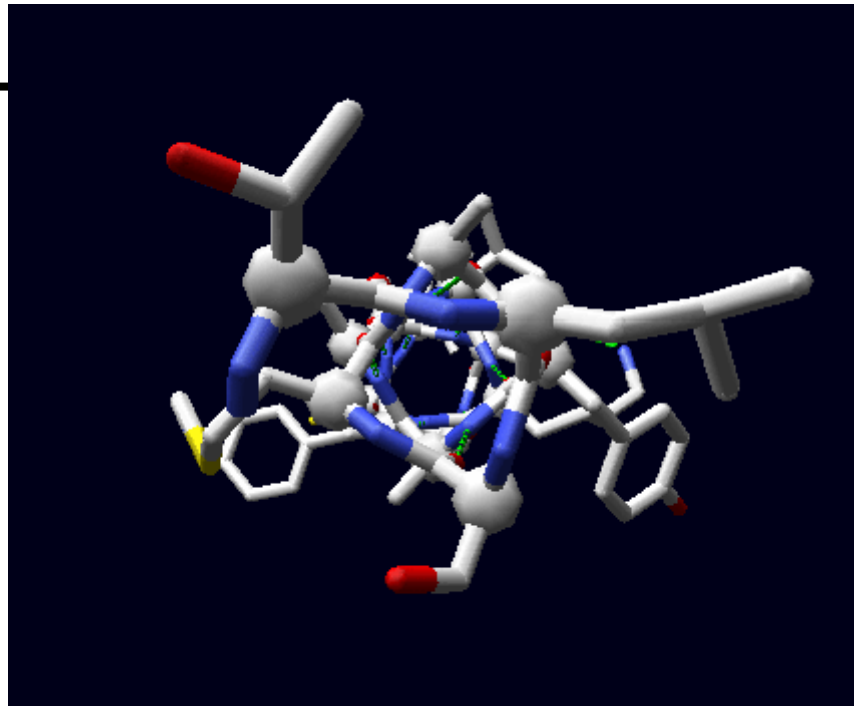


Physics of secondary structures

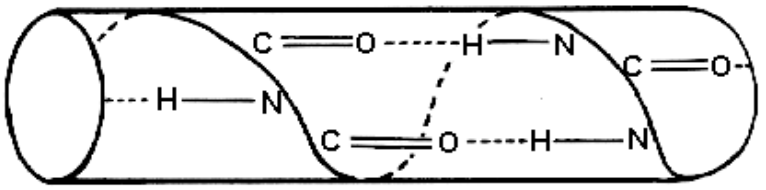
- Two main opposing forces
  - sidechain conformational entropy
  - mainchain hydrogen bonding.
- This predicts:
  - Helix propensity Ala>Leu>Ile>Val
- Other factors
  - Polarity (low helical propensity of Ser, Thr, Asp and Asn)

Secondary structures -Helix





Toilet roll representation of the main chain hydrogen bonding in an alpha-helix.

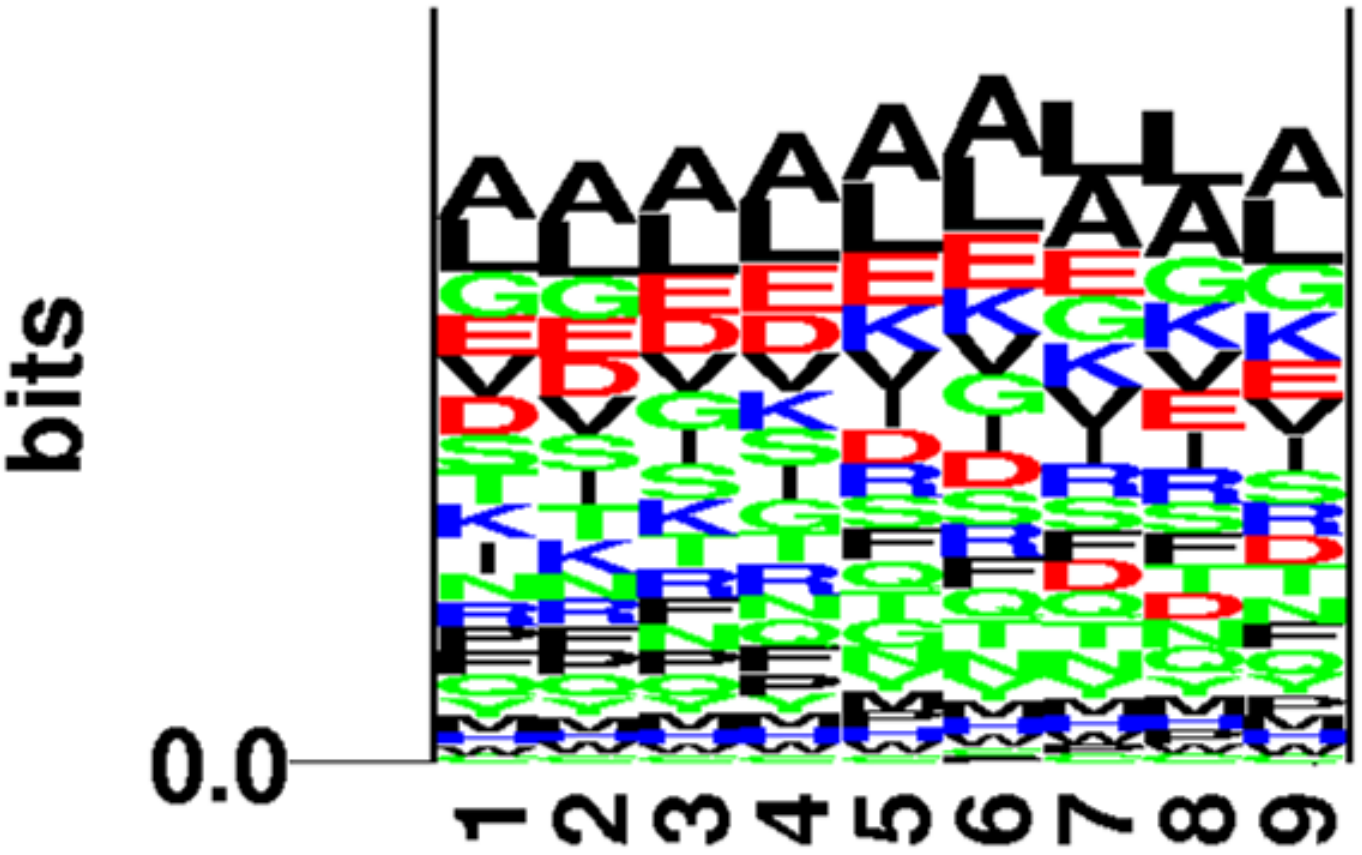


Amino terminus

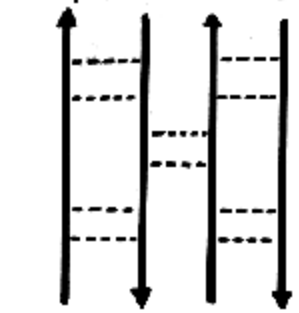
Carboxy terminus

Arne Elofsson (arne@bioinfo,se)

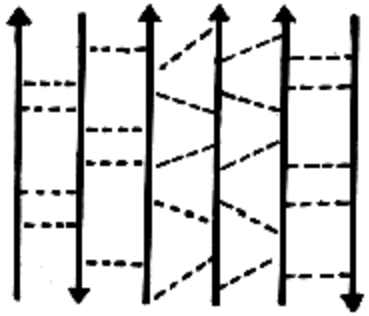
Amino acid preferences in α-Helix

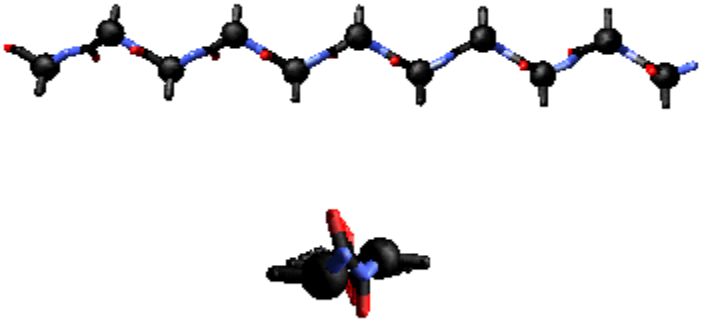


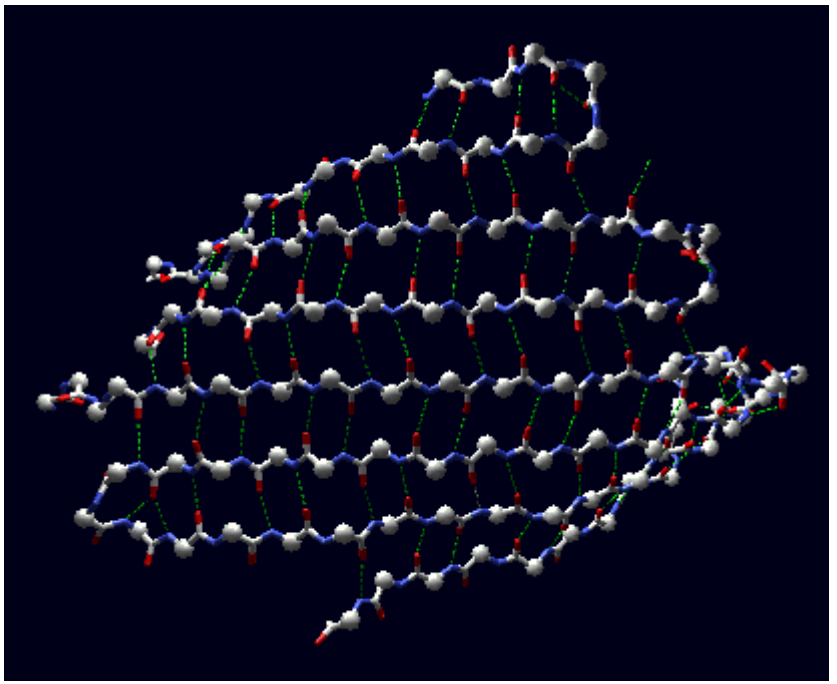
Secondary Structure - Sheet



The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.

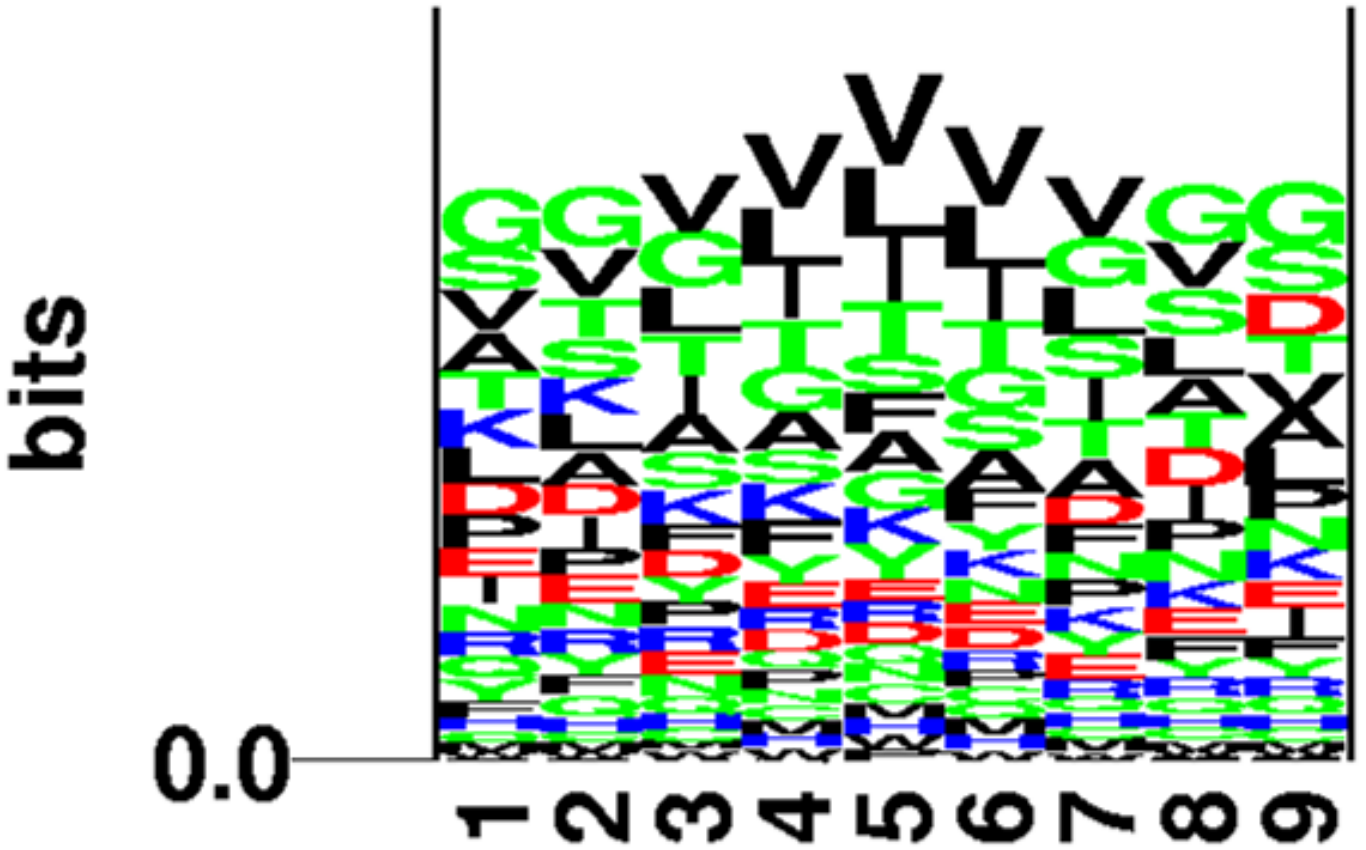




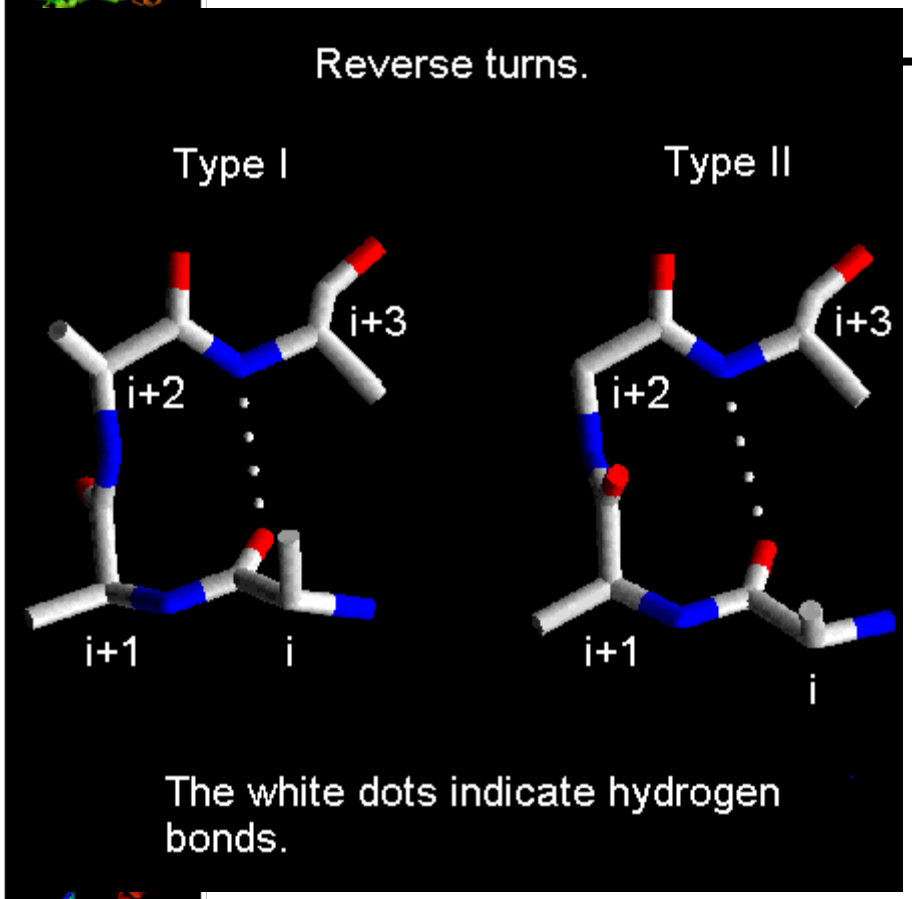


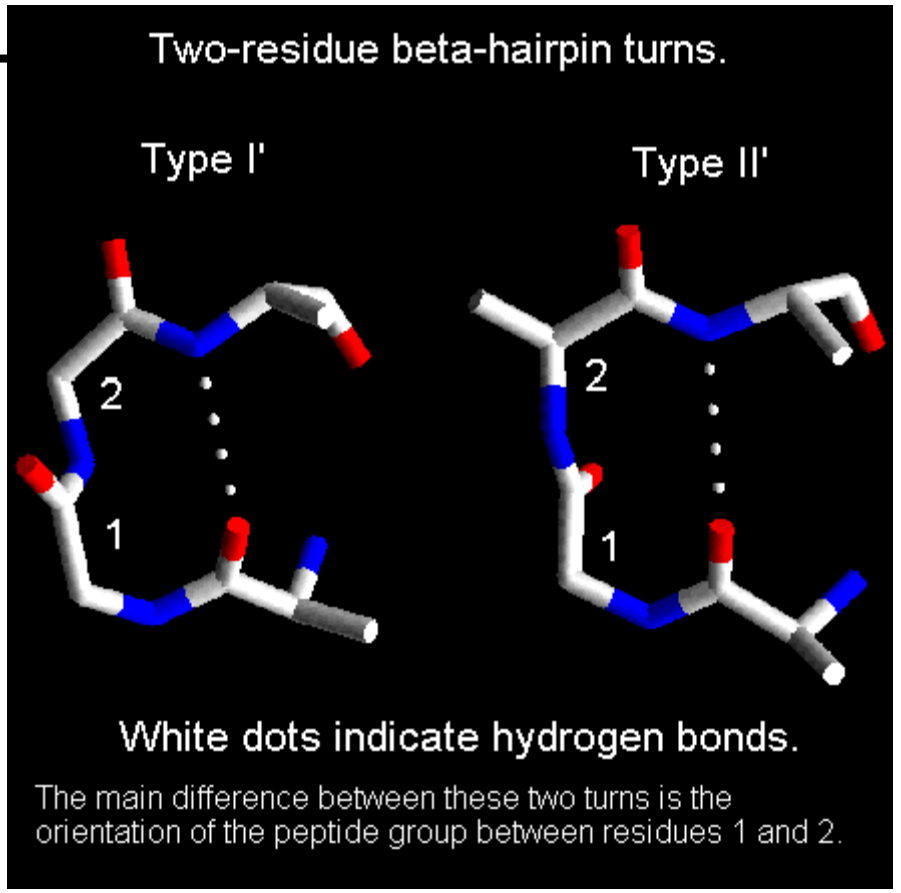
Arne Elofsson (arne@bioinfo,se)

Amino acid preferences in β-Strand

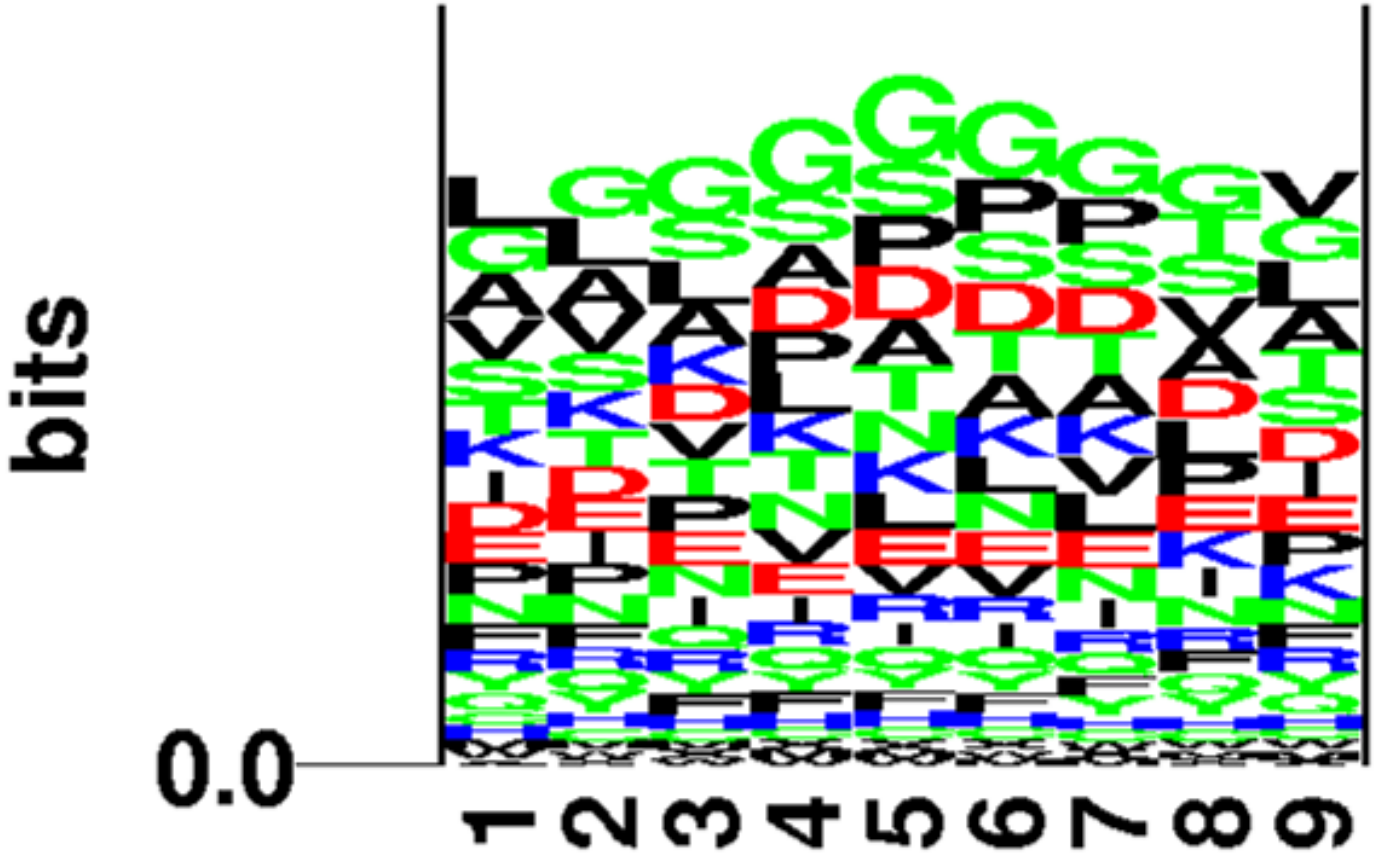


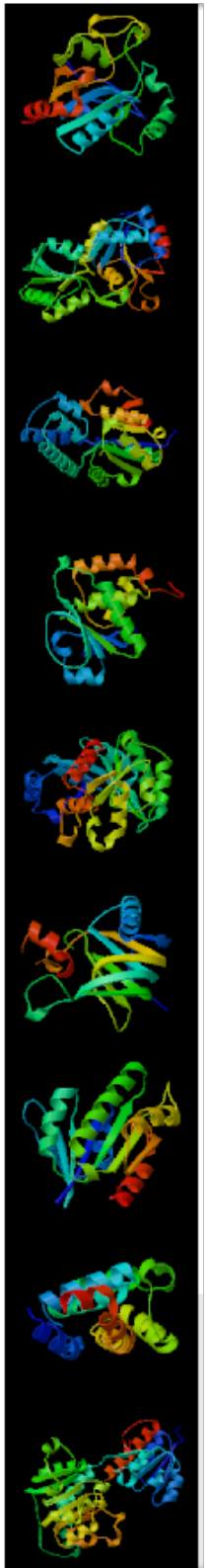
Secondary structure - turns





Amino acid preferences in coil



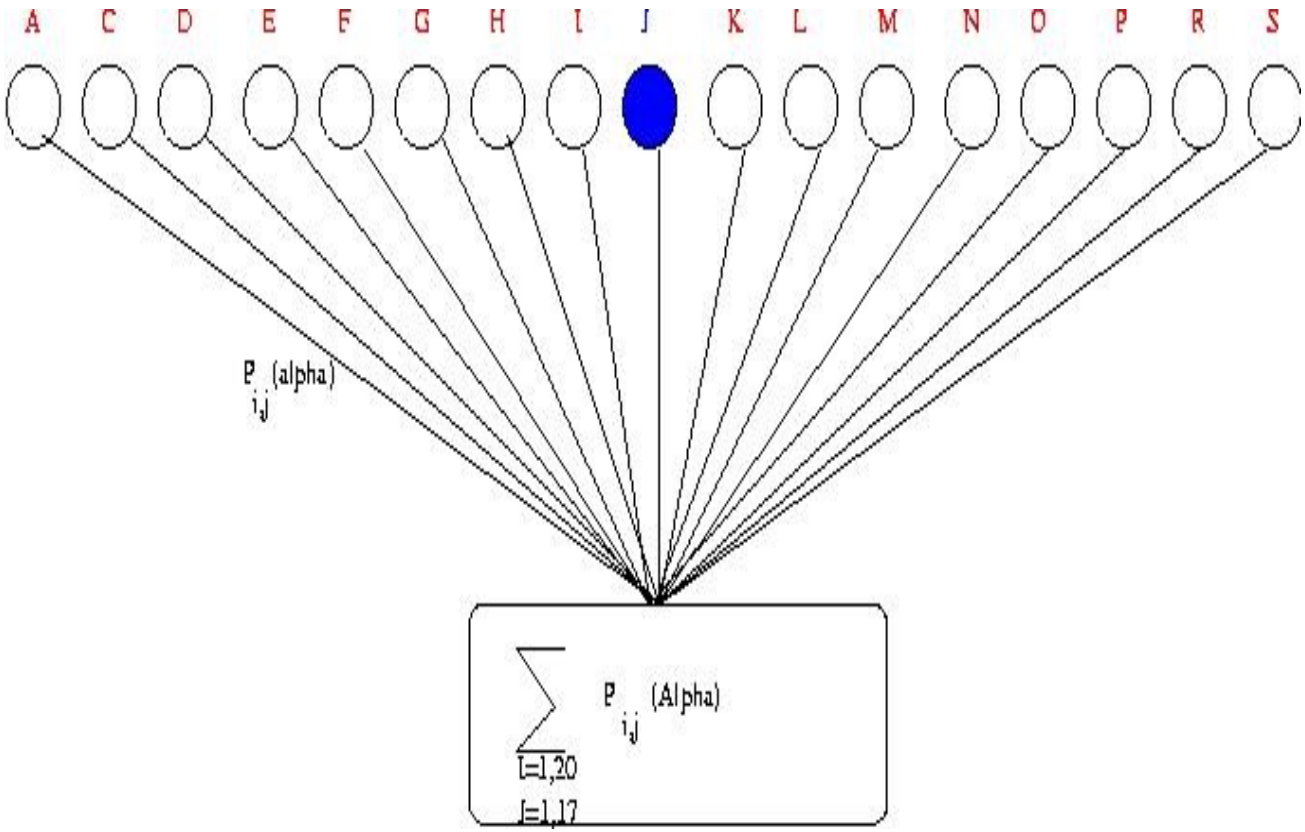


Secondary Structure Predictions

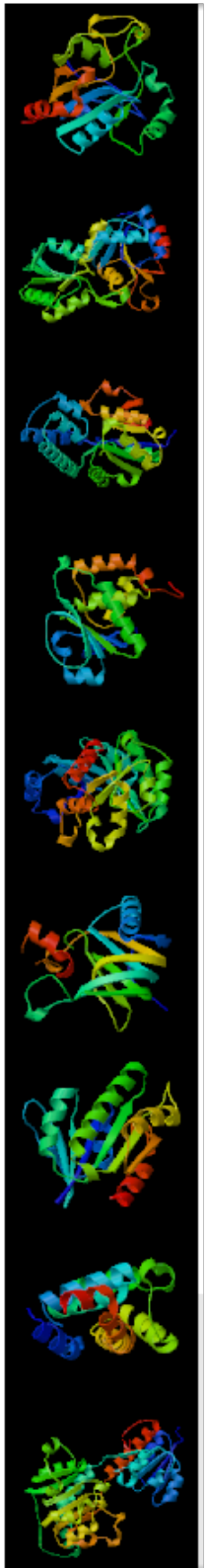
Some highlights in performance

- 1974 Chou and Fasman 50%
- 1978 GOR III 62%
- 1993 PhD 72%
- 2000 PsiPred 76%

Secondary structure prediction 1st generation methods



Arne Elofsson (arne@bioinfo,se)

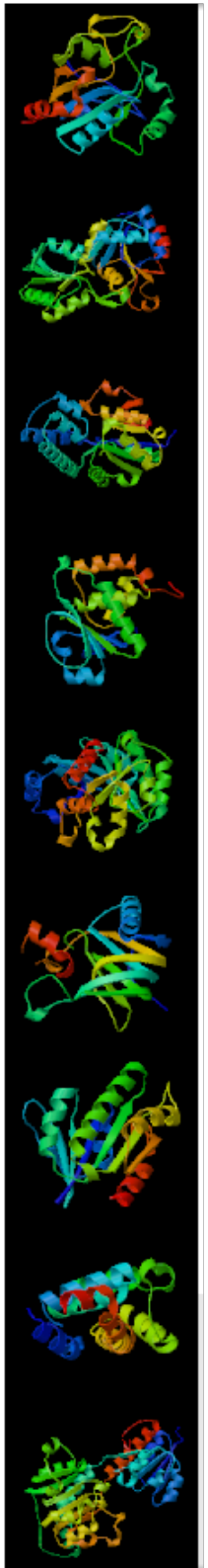


Chou and Fassman

Chou and Fassman

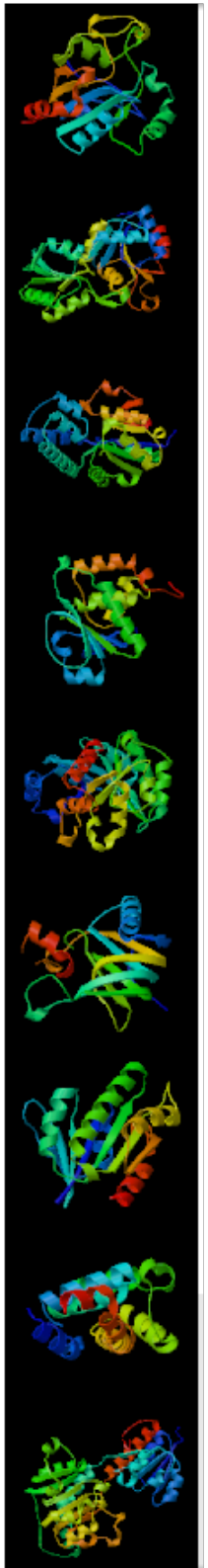
- Assign all residues the appropriate set of parameters.
- Scan through the peptide and identify helical regions
- Repeat this procedure to locate all of the helical regions in the sequence.
- Scan through the peptide and identify sheet regions.
- Solve conflicts between helical and sheet assignments
- Identify turns

- Claims of around 70-80% - actual accuracy about 50-60%



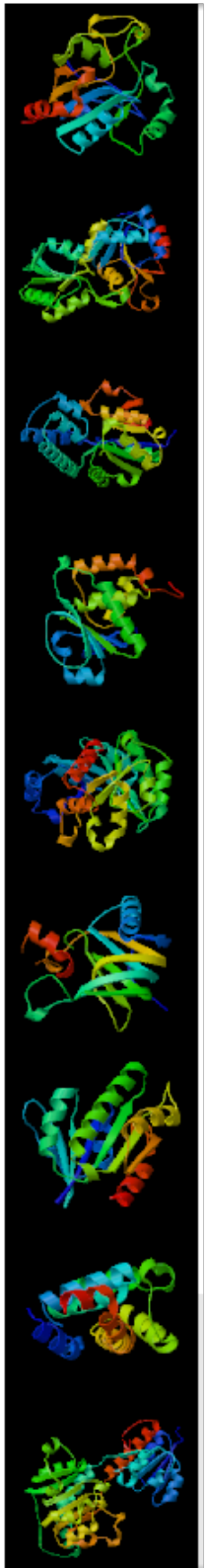
Chou-Fasman

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Ala	142	83	66	0.06	0.076	0.035	0.058
Arg	98	93	95	0.070	0.106	0.099	0.085
Asp	101	54	146	0.147	0.110	0.179	0.081
Asn	67	89	156	0.161	0.083	0.191	0.091
Cys	70	119	119	0.149	0.050	0.117	0.128
Glu	151	37	74	0.056	0.060	0.077	0.064
Gln	111	110	98	0.074	0.098	0.037	0.098
Gly	57	75	156	0.102	0.085	0.190	0.152
His	100	87	95	0.140	0.047	0.093	0.054
Ile	108	160	47	0.043	0.034	0.013	0.056
Leu	121	130	59	0.061	0.025	0.036	0.070
Lys	114	74	101	0.055	0.115	0.072	0.095
Met	145	105	60	0.068	0.082	0.014	0.055
Phe	113	138	60	0.059	0.041	0.065	0.065
Pro	57	55	152	0.102	0.301	0.034	0.068
Ser	77	75	143	0.120	0.139	0.125	0.106
Thr	83	119	96	0.086	0.108	0.065	0.079
Trp	108	137	96	0.077	0.013	0.064	0.167
Tyr	69	147	114	0.082	0.065	0.114	0.125
Val	106	170	50	0.062	0.048	0.028	0.053



Chou-Fasman

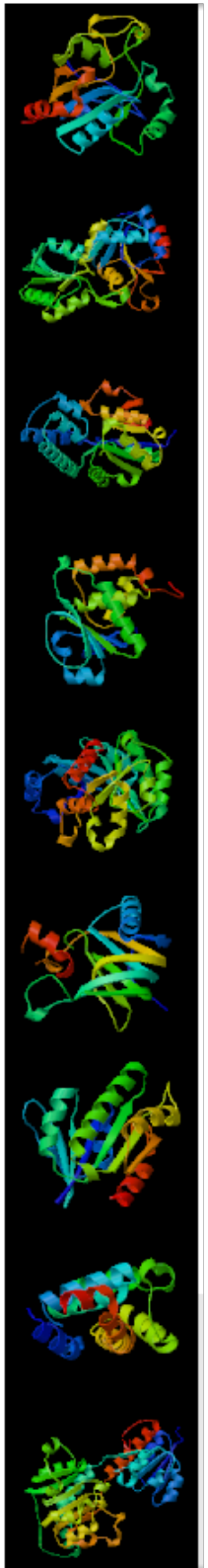
- General applicable
- Works for sequences with no solved homologs
- But the accuracy is low!



GOR III

Garnier, Osguthorpe, Robson, 1990

- Secondary structure depends on aminoacids propensities
  - As in Chou Fassman
- Also influences by neighboring residues
  - Helix capping
  - Turns etc
- How to include distant information.
- Performance approximately 67%



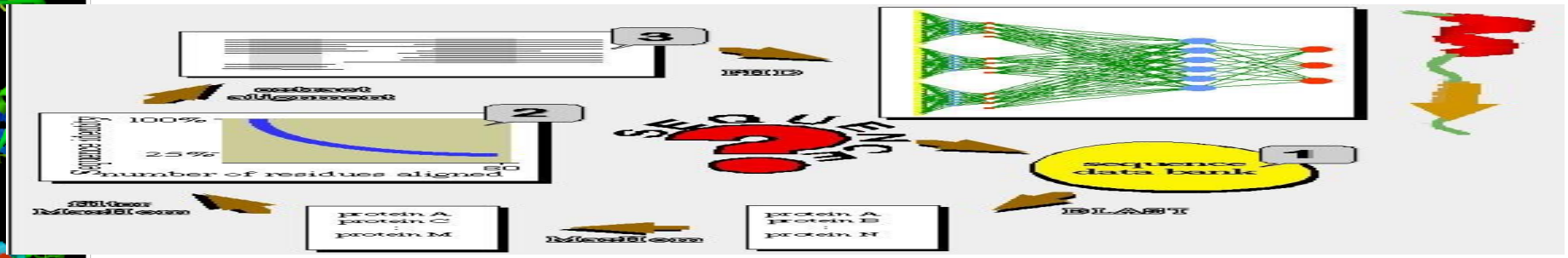
GOR III

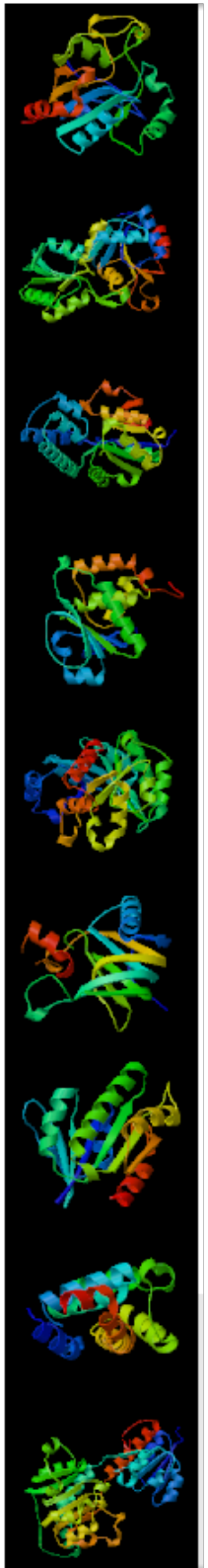
Garnier, Osguthorpe, Robson, 1990

The helix propensity tables thus have 20x17 entries.
Assign the state with the highest propensity

Status of predictions in 1990

- Too short secondary structure segments
- About 65% accuracy
- Worse for Beta-strands
- Example:

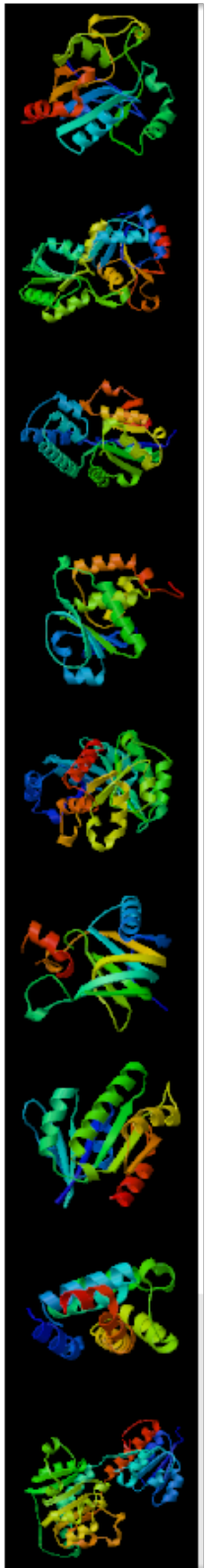




Secondary structure prediction 2nd generation methods

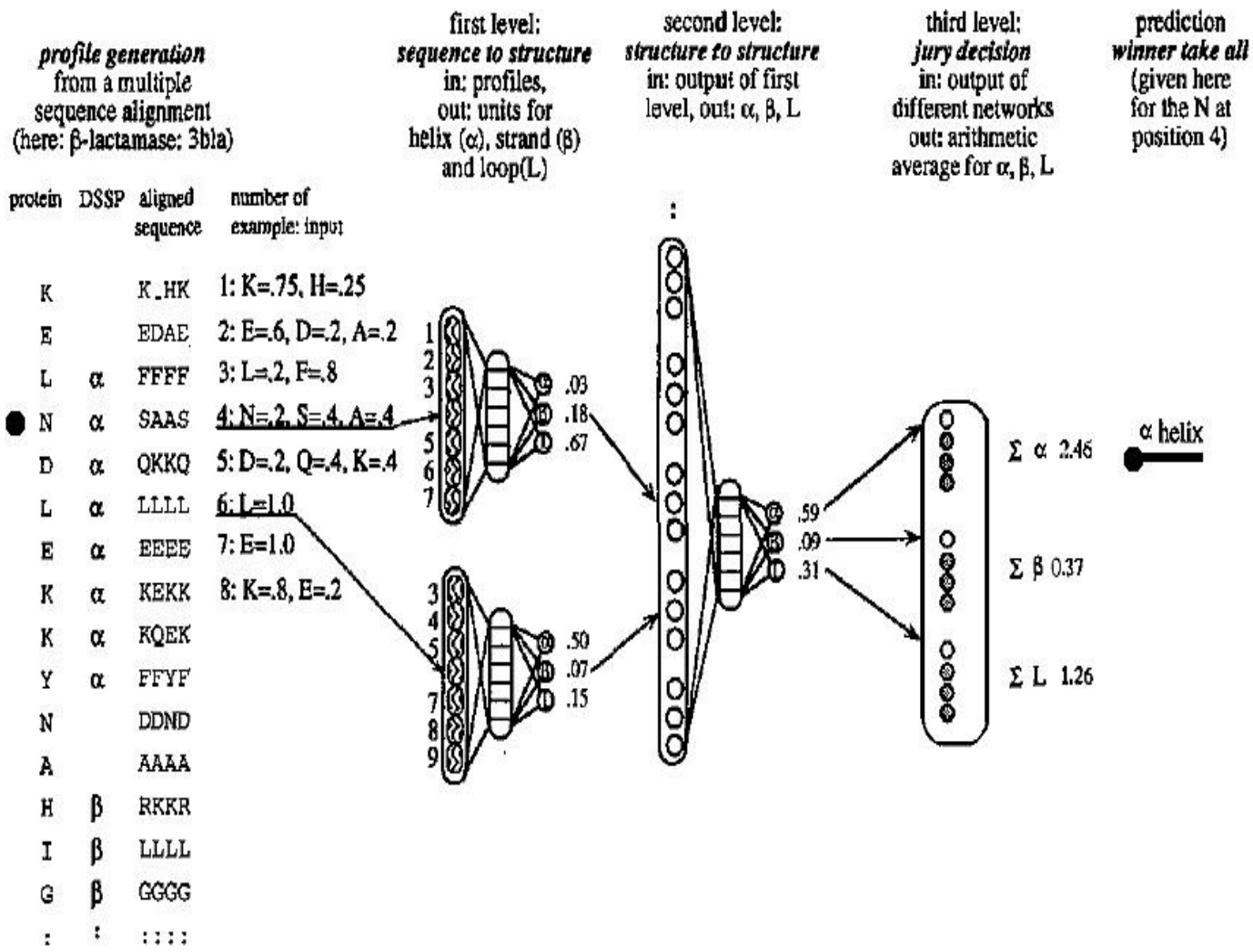
- sequence-to-structure relationship modelled using more complex statistics, e.g. artificial neural networks (NNs) or hidden Markov models (HMMs)
- evolutionary information included (profiles)
- prediction accuracy >70% (PhD, Rost 1993)

PhD (Rost & Sander, 1994)



Arne Elofsson (arne@bioinfo,se)

PhD-Input



Arne Elofsson (arne@bioinfo,se)

PhD-architecture

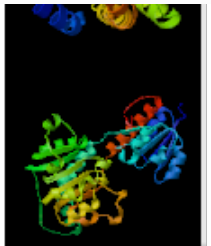
AUTHOR M. NOBLE, R. FAUPTIT, A. MUSACCHIO, M. SARASTE

```
      ....,....1....,....2....,....3....,....4....,....5....,..
AA    KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
OBS     EEEE          E--E    EEEEEE    EEEEEE  EEEEEEHHHEEEE

C+F   HHHHHHH         HHHHHH  EEEEEE    HHHHHH  EEEEEEHHHHHHH   59%
GOR   HHHHHHHH        HHHH    EEEEEE    EEEEHH   HHH   HHHHHHH   65%

PHD     EEEEEE        EEE   EEEEEEEE    HHHHHH    EEEE HHEEEE   72%
Rel   948999972587775211443884899847697314344045955111321221558
      *  ******  ******      **  ****  ****        ****      ***
```

Burkhard Rost (Columbia New York)

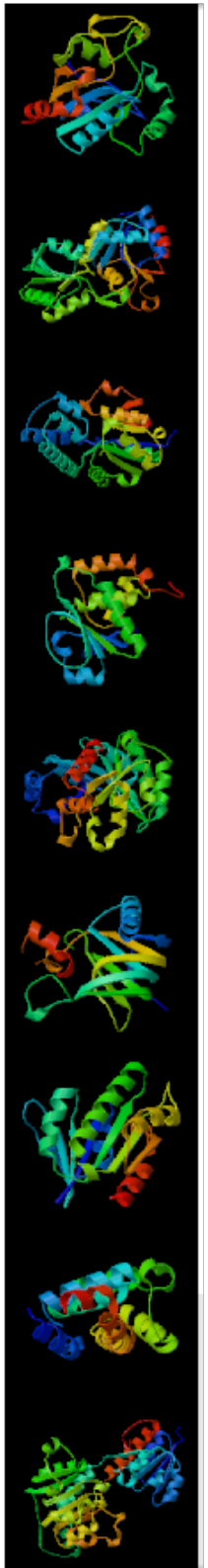


PhD-predictions

Genome Reviews

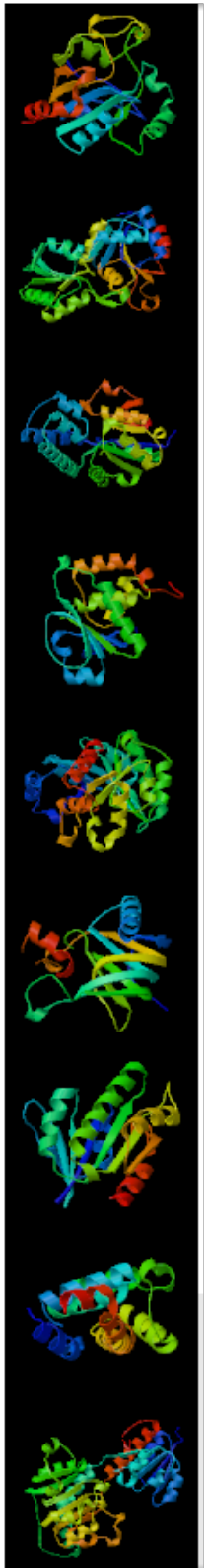
Genome Reviews: Release Stats

Release stats	
Release version	20
Release date	07 February 2005
Number of entries:	353
Number of complete genomes	207
Number of nucleotides	674,225,858
Number of protein coding sequences	625,623



PhD summary

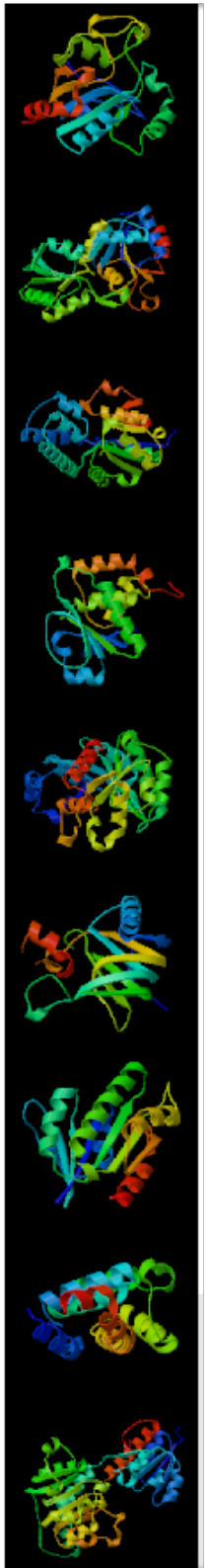
- First methods with >70% Q3
- Correct length distributions
- Much better beta strand predictions
- Good correlation between score and accuracy
- Better predictions for larger multiple sequence alignments



Best method !

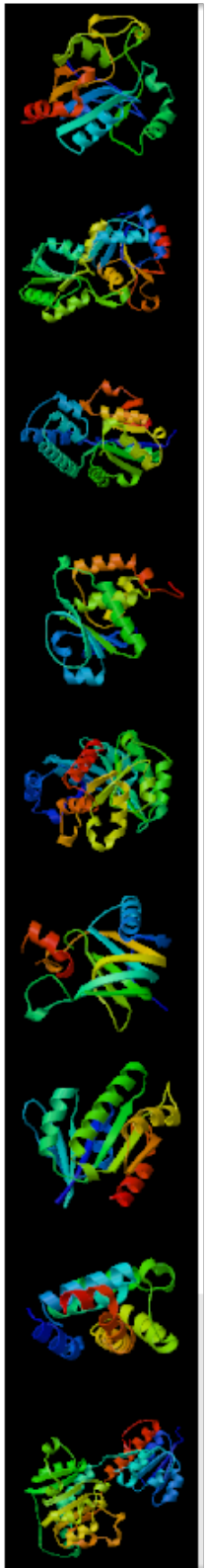
- Secondary structure ``prediction'' by homology

– If sequence of unknown secondary structure has a homologue of known structure, it is *more accurate* to make an alignment and *copy the known secondary structure* over to the unknown sequence, than to do ``ab initio'' secondary structure prediction.



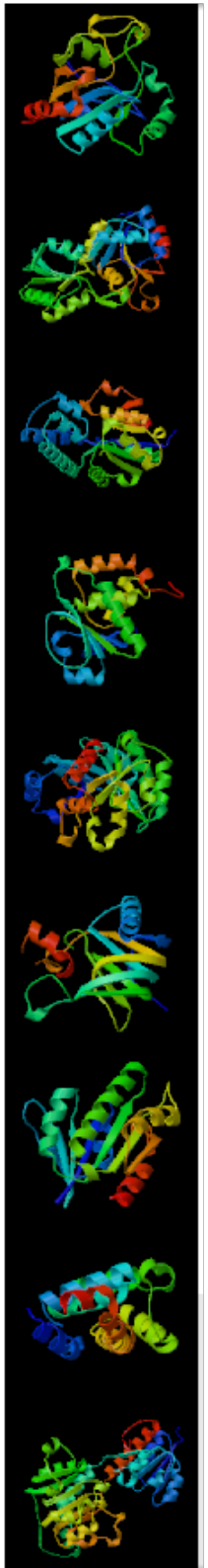
Nearest neighbour methods

- Generate fragment of proteins with known structures
- Align sequence to all these
- Calculate the “average” secondary structure of aligned residues
- Filter
- Prediction accuracy > 70%
- Not sustained in CASP ?



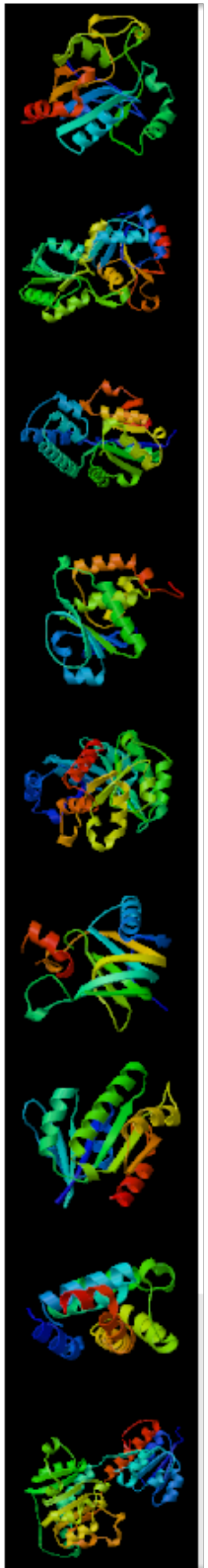
3rd generation methods

- enhanced evolutionary sequence information (PSI-BLAST profiles) and larger sequence databases takes Q3 to > 75%
- PHD and PSIPRED are the best known methods



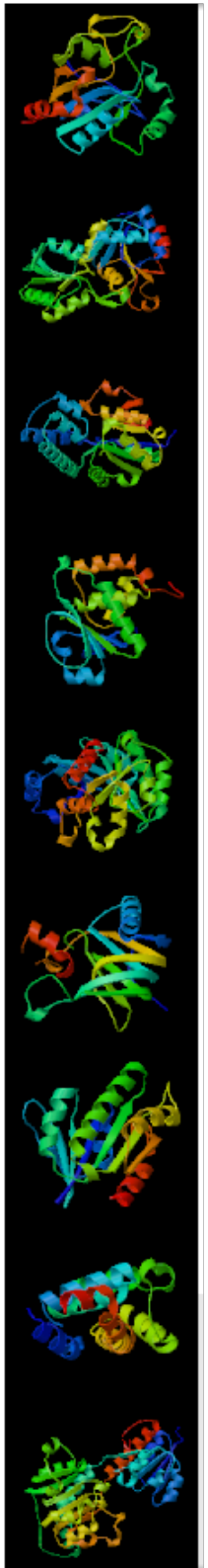
PSIPRED

- Similar to PhD
- PSIBLAST to detect more remote homologs
- only two layers
- SVM or ANN gives similar performance



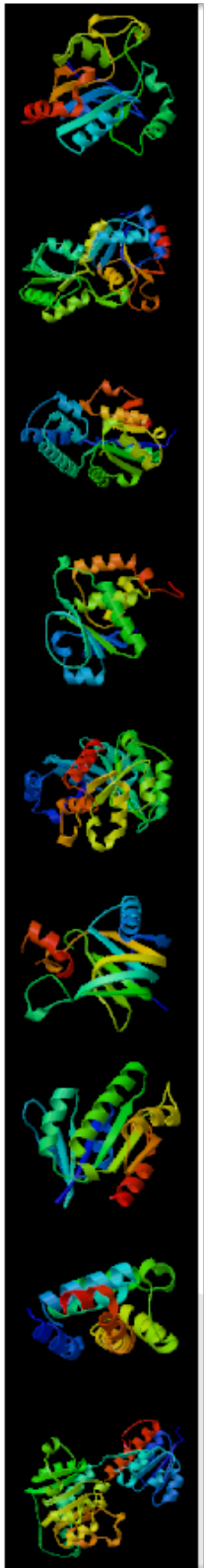
Current Status of Secondary Structure predictions

- Best Methods
  - PsiPred
  - Sam-T02
  - Prof
- About 75%-76% accuracy
- Improvement mainly due to:
  - Larger Databases
  - PSI-BLAST



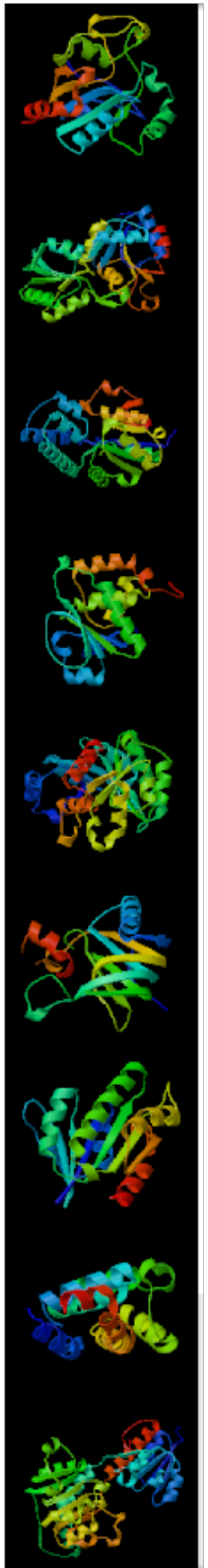
Other secondary structure prediction methods

- turn prediction
- transmembrane helix prediction
- coiled coil
- Dissorder predictions
- contact prediction, disulphides



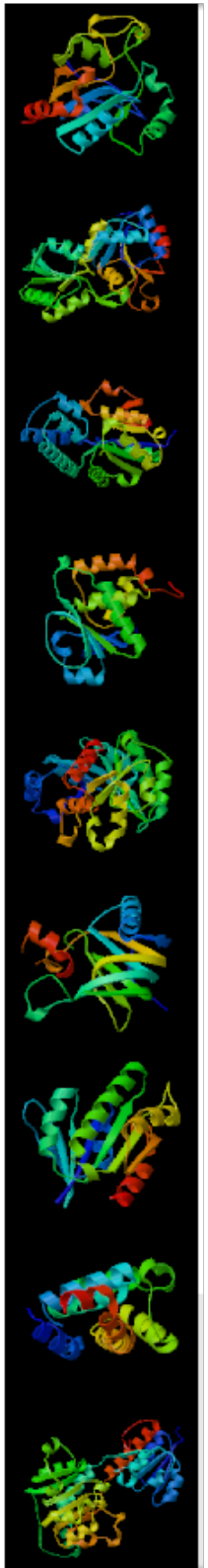
What use is it?

- No 3D means no clues to detailed function, so...
- Accurate secondary structure predictions help sequence analysis: finding homologues, aligning homologues, identifying domain boundaries.
- Can help true 3D prediction



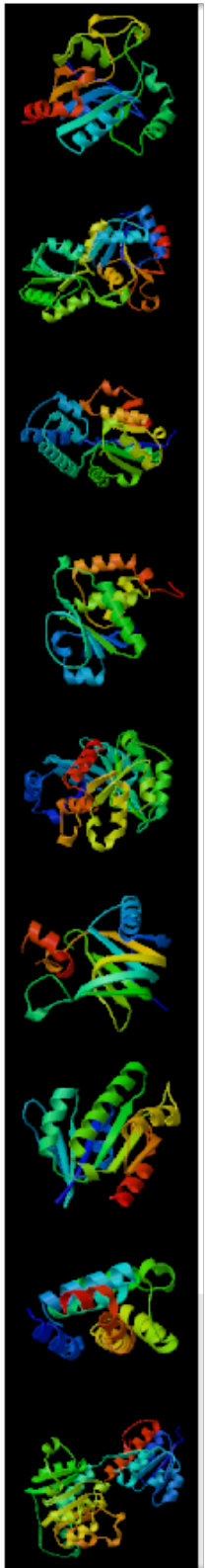
Future improvements to SSP

- Long range information
  - Baker
- Folding pathway and/or 3D-information
- HMMSTR and I-Sites



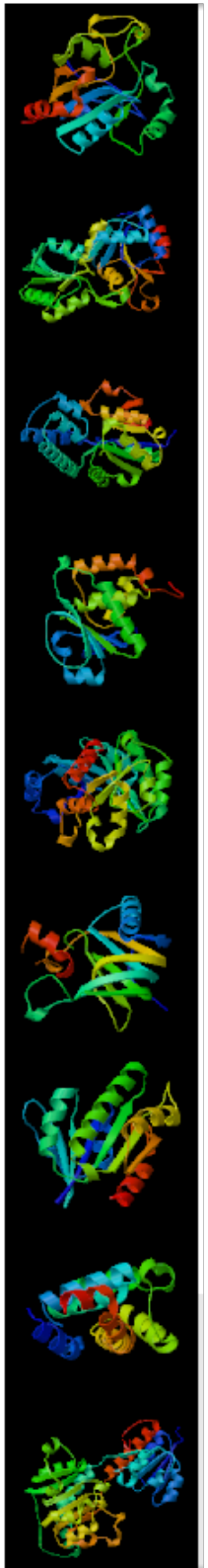
Why protein modeling?

- Experimental effort to determine protein structure is very large and costly
- The gap between the size of the protein sequence data and protein structure data is large and increasing
- Close to 50% of all new sequences can be homology modeled



Why do we need structural models?

- only 20% of all proteins have a homologue in PDB
- for ~ 70% of the proteins a suitable structure from which to build a 3D model is available.
- predict functions of proteins that share low degrees of sequence similarity
- identify proteins that may have new folds



Scope of the Problem

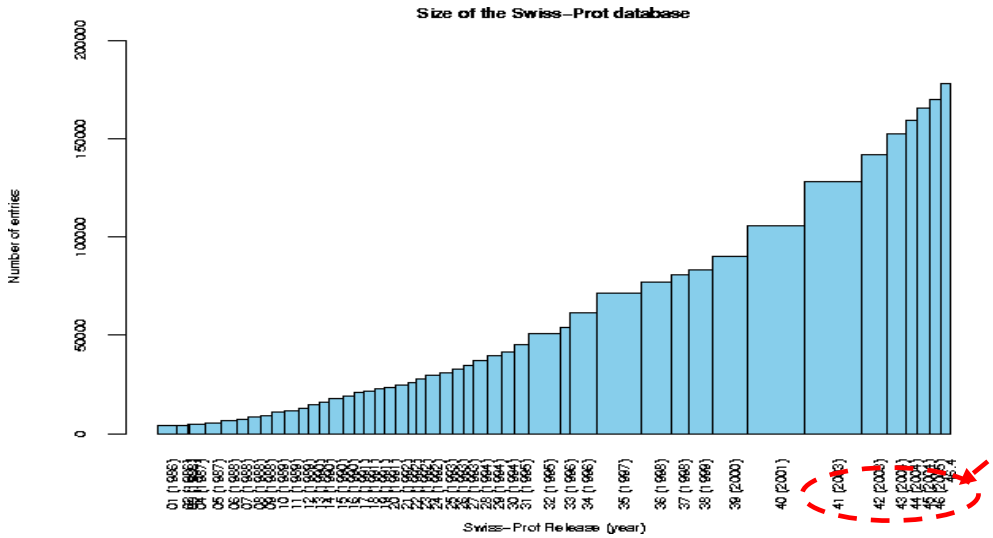
- ~90% of *new* globular proteins share similar folds with known structures, implying the general applicability of comparative modeling methods for structure prediction
- general applicability of template-based modeling methods for structure prediction (currently 60-70% of new proteins, and this number is growing as more structures being solved)
- NIH *Structural Genomics Initiative* plans to experimentally solve ~10,000 “unique” structures and predict the rest using computational methods

Why do we need homology modeling ?

PDB Holdings List: 15-Feb-2005

		Proteins, Peptides, and Viruses
Exp.	X-ray Diffraction and other	23352
Tech.	NMR	3626
	Total	26978

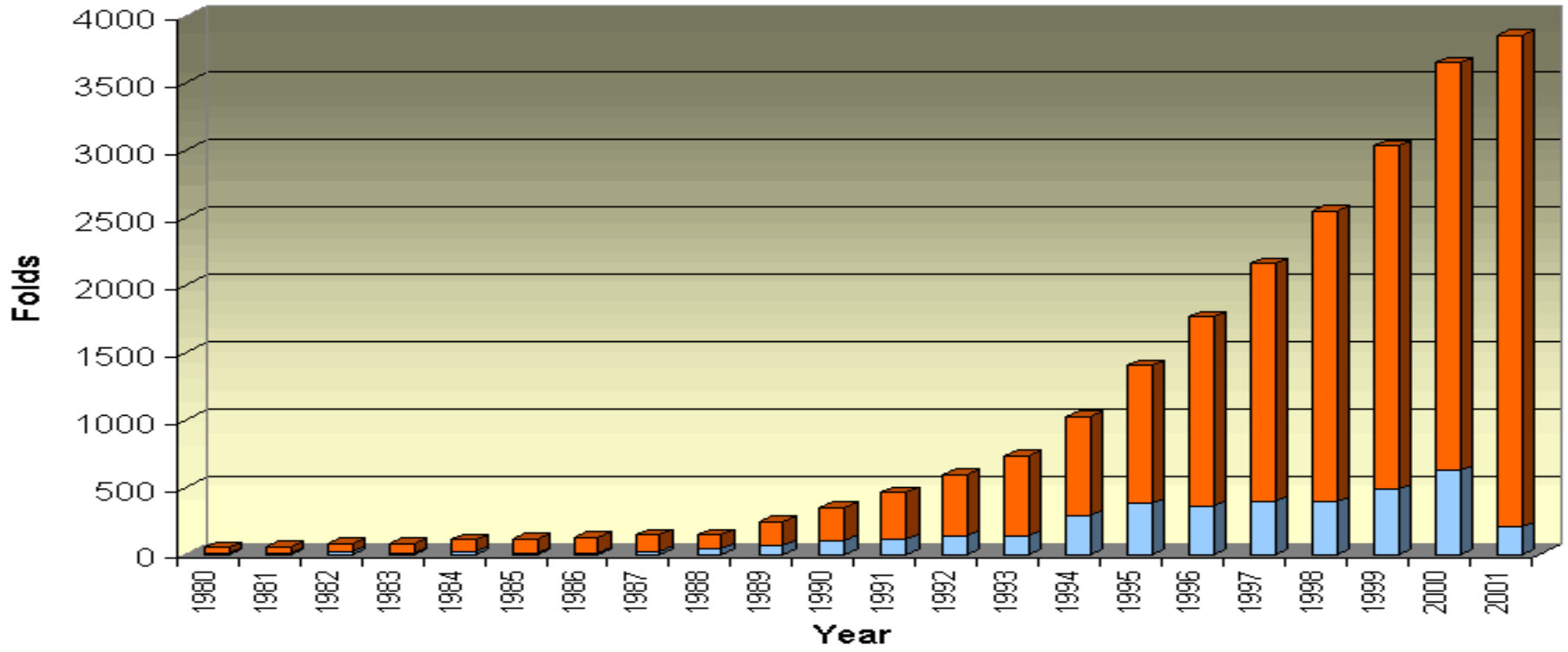
To be compared with:



Arne Elofsson (arne@bioinfo,se)

How many structures are there ?

Protein Data Bank (PDB) Status: March 12, 2002

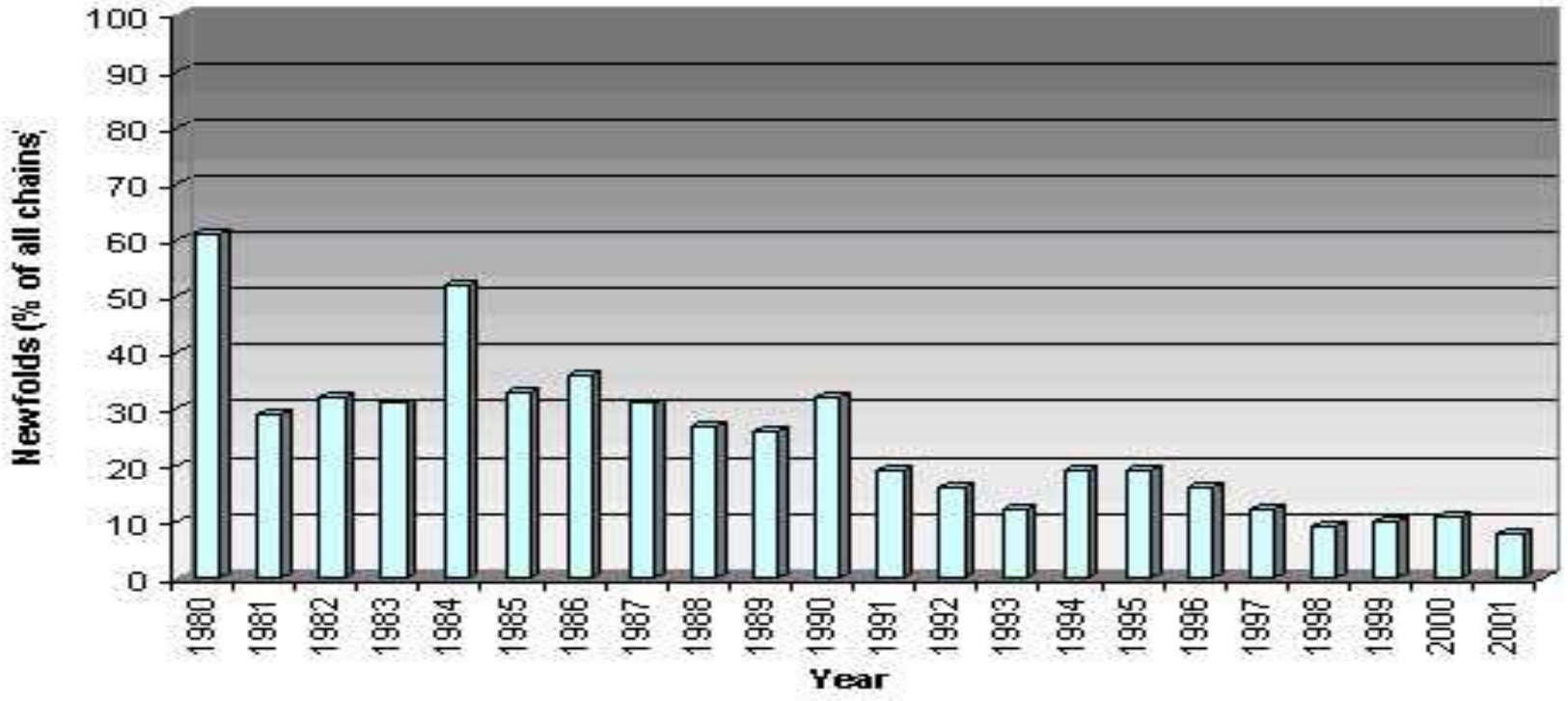


Source: http://www.rcsb.org/pdb/holdings.html

Arne Elofsson (arne@bioinfo,se)

How many folds are there ?

Structural Classification of Proteins (SCOP):
Status (1 Mar 2002) based on 13220 PDB entries



Source: http://scop.berkeley.edu/count.html

Arne Elofsson (arne@bioinfo,se)

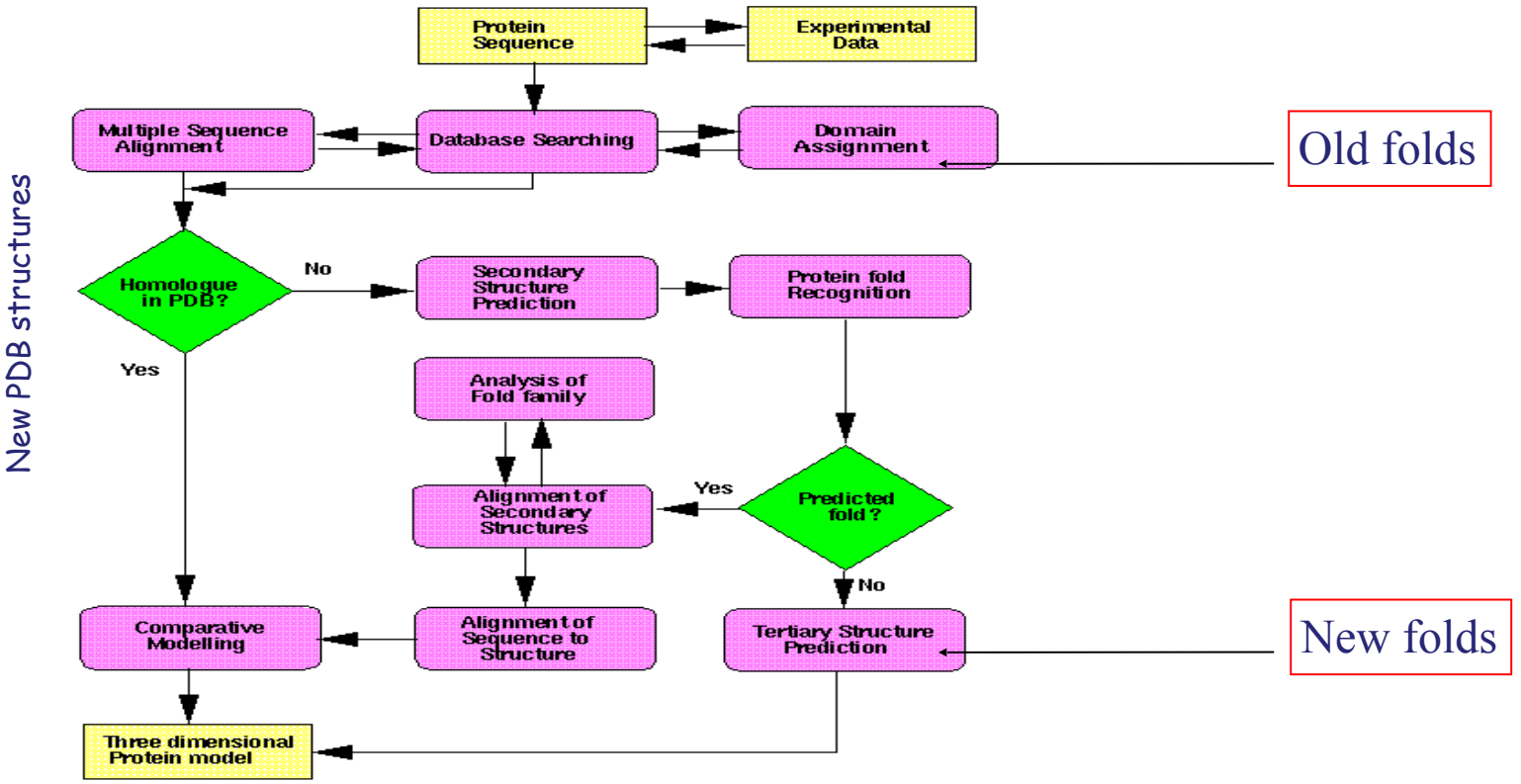
Swiss-Prot database

Probabilities of SWISS-MODEL accuracy for target-template identity classes.

Percent sequence identity[a]	Total number of models[b]	Percent[c] models with rmsd lower than 1 Å	Percent models with rmsd lower than 2 Å	Percent models with rmsd lower than 3 Å	Percent models with rmsd lower than 4 Å	Percent models with rmsd lower than 5 Å	Percent models with rmsd higher than 5 Å
25-29	125	0	10	30	46	67	33
30-39	222	0	18	45	66	77	23
40-49	156	9	44	63	78	91	9
50-59	155	18	55	79	86	91	9
60-69	145	38	72	85	91	92	8
70-79	137	42	71	82	85	88	12
80-89	173	45	79	86	94	95	5
90-95	88	59	78	83	86	91	9

a: Range of sequence identity between target and template sequence.
b: Total number of models in any given class of sequence identity. The table summarises 1201 model - control structure pairs.
c: Probability in percent that a model, sharing X% sequence identity with its template, deviates by 1 Å or less from the corresponding experimental control structure. The following columns provide these probabilities for other rms deviations.

PDB New Fold Growth

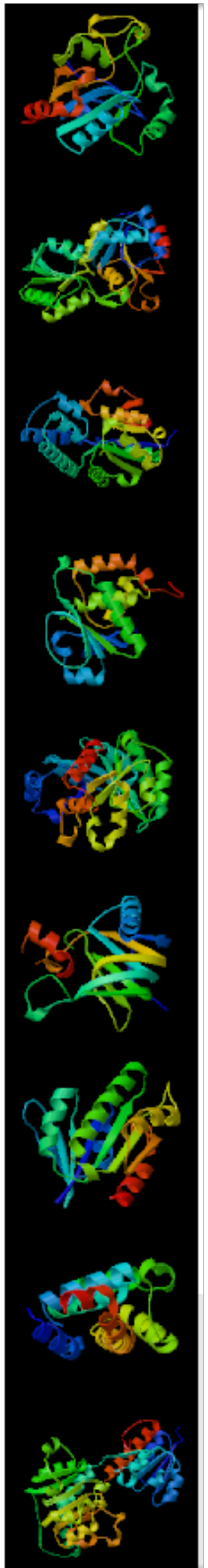


Arne Elofsson (arne@bioinfo,se)

Identification of new folds

Source: http://www.rcsb.org/pdb/holdings.html

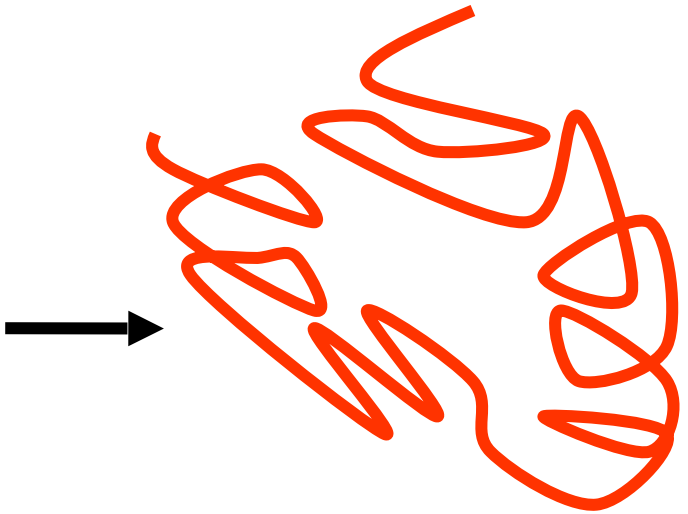
Arne Elofsson (arne@bioinfo,se)

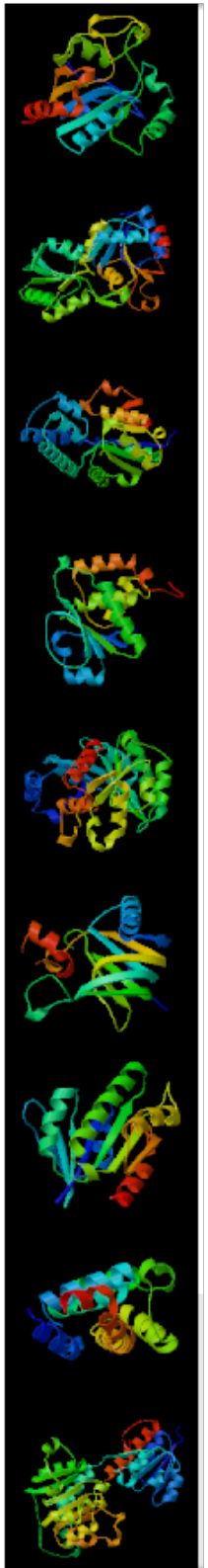


Homology Modeling

- Given a sequence what is the best way of mounting it onto a known structure

GHIKLSYTVNEQN
LKPERFFYTSAVAIL





What is the basis for homology modelling?

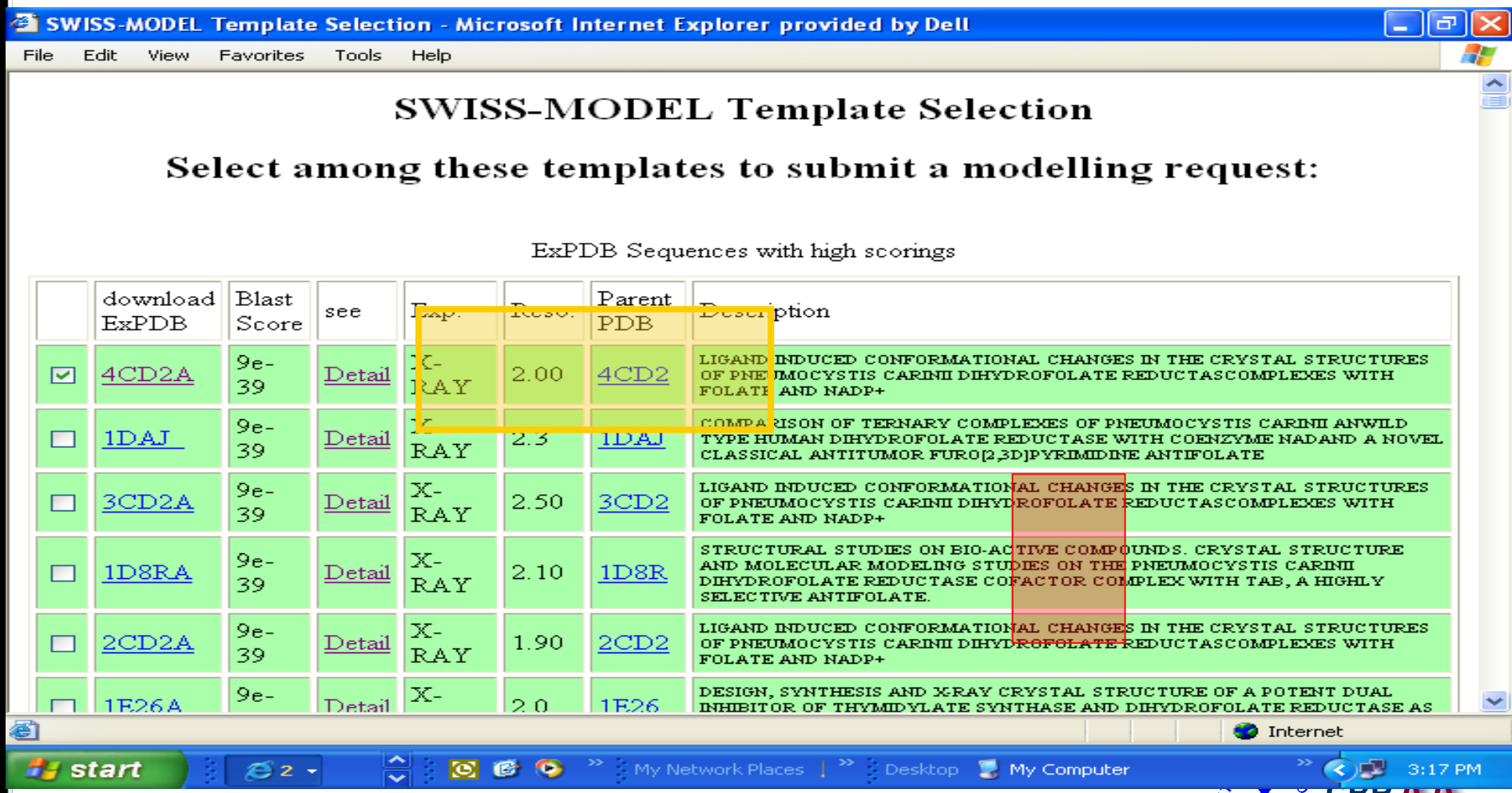
- The relative RMSD of the α-carbon coordinates is ~ 1 if the protein core share 50% identity.
- Protein sequences with > 70% similarity allow construction of models with < 3 Å RMSD
- Reduction to:

- Loop structure modeling (connections αα, ββ, αβ, βα
- Side-chain modeling (energy refinement)

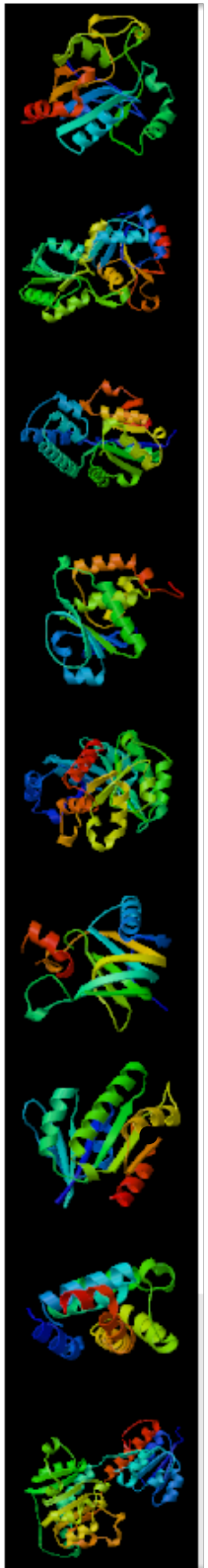
Arne Elofsson (arne@bioinfo,se)

Model accuracy. Swiss-model.

1200 models sharing 25-95% sequence identity with the submitted sequences
(www.expasy.ch/swissmod)



Arne Elofsson (arne@bioinfo,se)



Input requirements for Homology Modelling

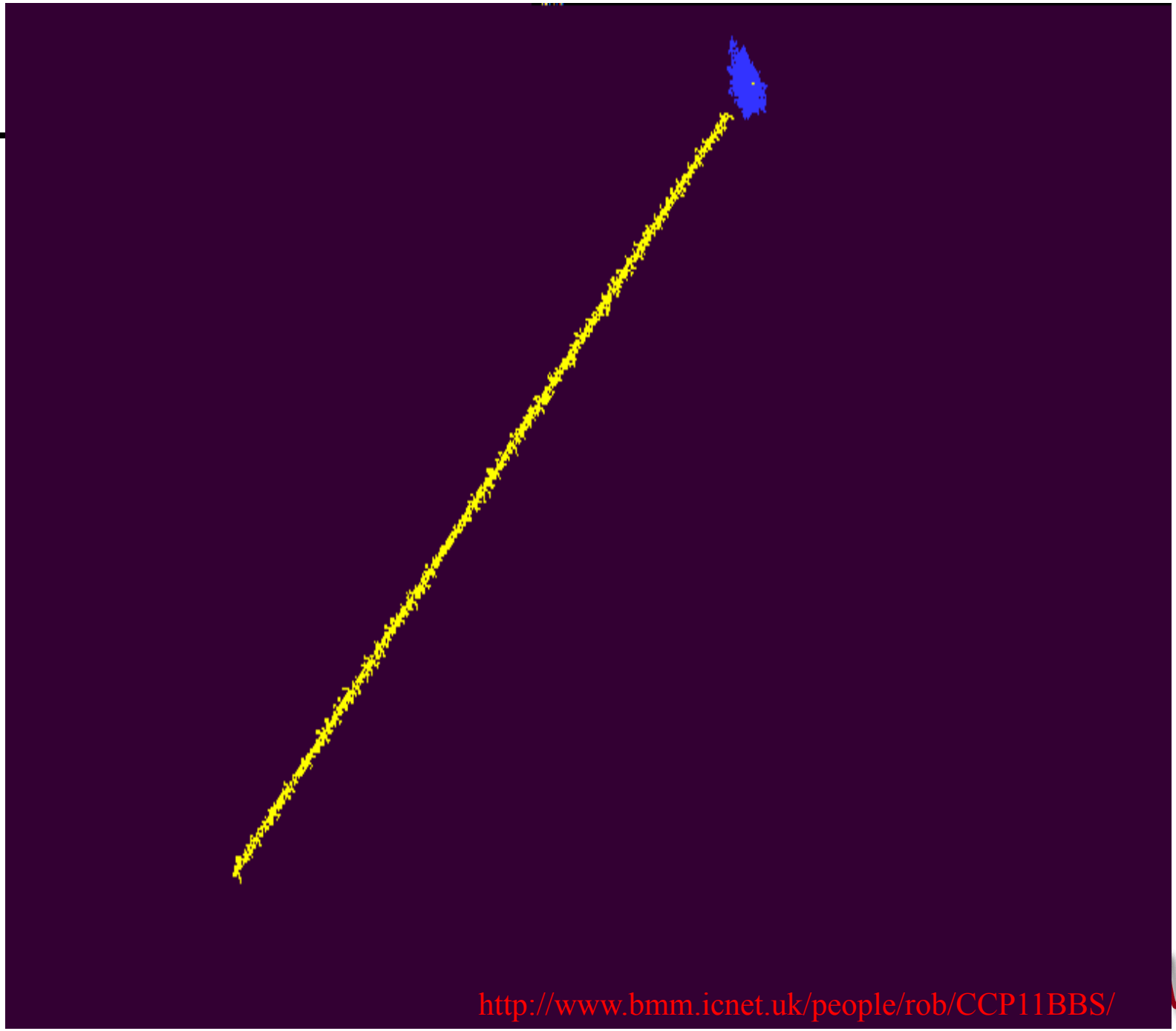
TARGET SEQUENCE (primary protein sequence with unknown structure)

TEMPLATE (protein whose 3D structure has already been determined)

SEQUENCE ALIGNMENT (using Clustal W) between template and target sequence

Arne Elofsson (arne@bioinfo,se)

Prediction

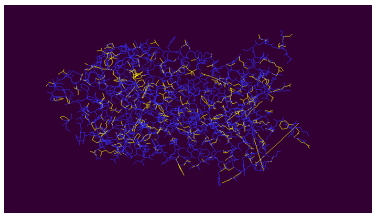


http://www.bmm.icnet.uk/people/rob/CCP11BBS/

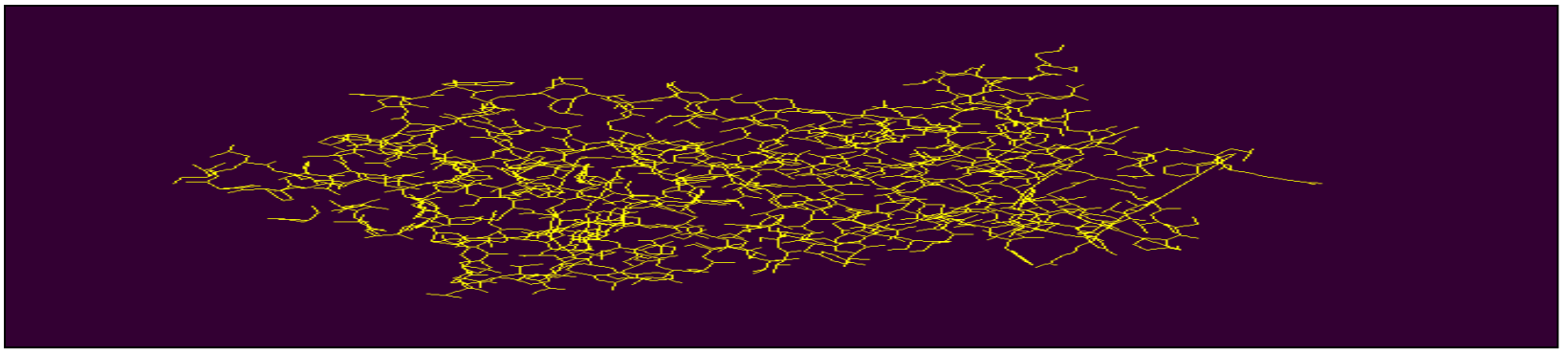
Find the appropriate template

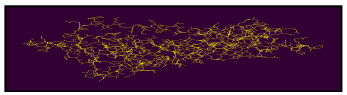
SWISS-MODEL Blast

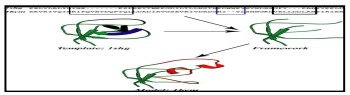
Find the Appropriate Modelling Template(s)



Please enter your sequence in **FASTA** format.

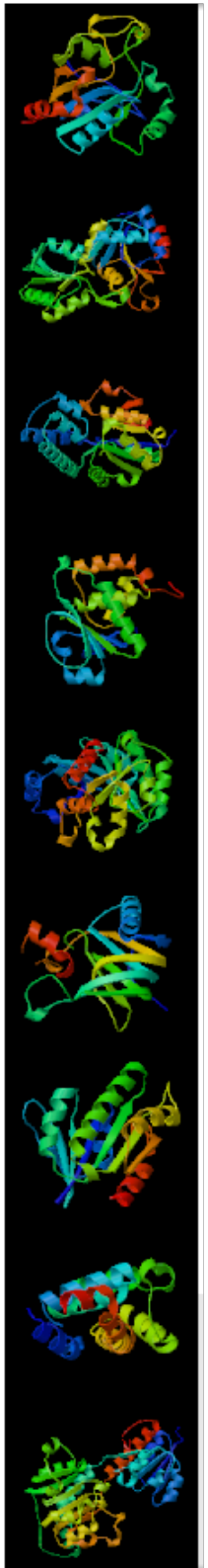




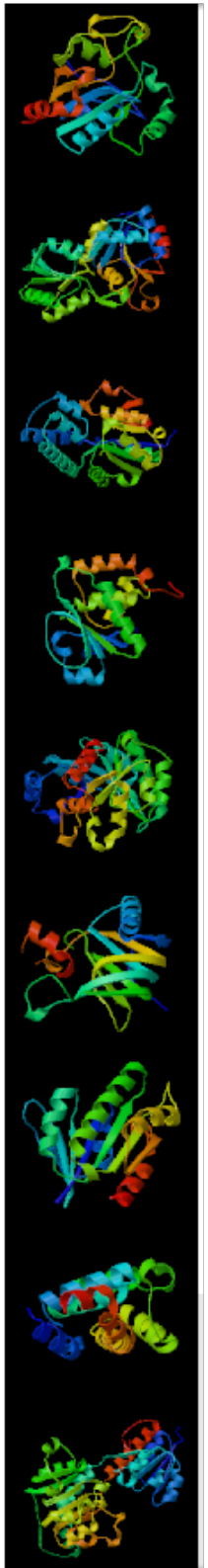


Source: http://www.expasy.org/swissmod/SM_Blast.html

Arne Elofsson (arne@bioinfo,se)



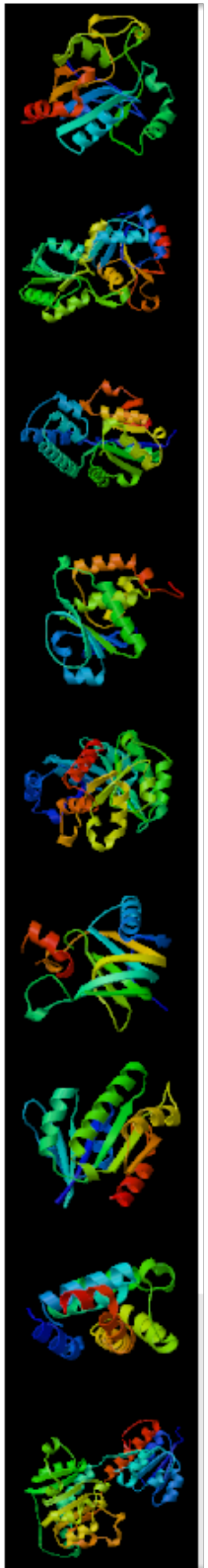
Choose a template



Template search results

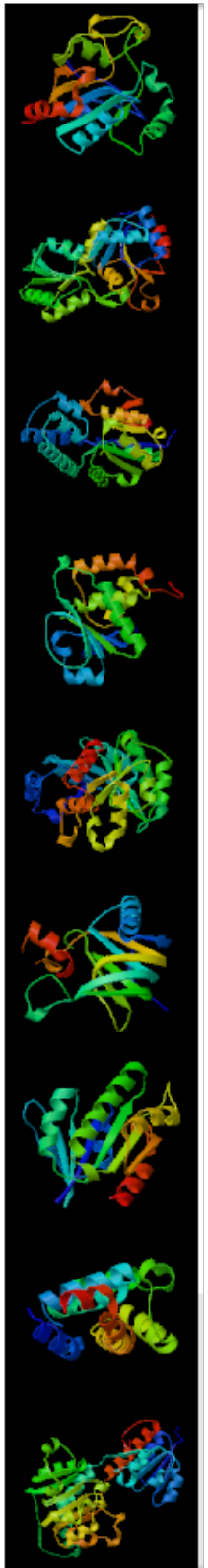
4CD2A top
LIGAND INDUCED CONFORMATIONAL CHANGES IN THE CRYSTAL
STRUCTURES OF PNEUMOCYSTIS CARINII DIHYDROFOLATE REDUCTAS
COMPLEXES WITH FOLATE AND NADP+
MOL_ID: 1;
MOLECULE: DIHYDROFOLATE REDUCTASE;
CHAIN: A;
SYNONYM: PCDHFR;
EC: 1.5.1.3;
ENGINEERED: YES
MOL_ID: 1;
ORGANISM_SCIENTIFIC: PNEUMOCYSTIS CARINII;
ORGANISM_COMMON: BACTERIA;
V.CODY,N.GALITSKY,D.RAK,J.R.LUFT,W.PANGBORN,S.F.QUEENER
Length = 202

Score = 157 bits (393), Expect = 9e-39
Identities = 82/220 (37 Positives = 138/220 (62 Gaps = 22/220 (10
Query: 232
RDLTMIVAVSSPNLGIGKKNSMPWHIKQEMAYFANVTSSTESSGQLEEGKSKIMNVVIMG 291
LT IVA GIG NS PW K E YF VTS E MNVV
MG
Sbjct: 1 KSLTLIVALTT-SYGIGRSNSLPWKLKKEISYFKRVTSFVPTFDSFES-----
MNVVLMG 54

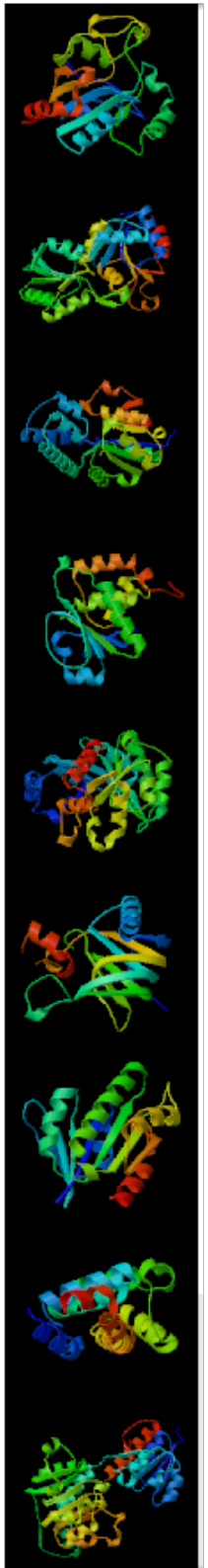


Mounting the sequence onto the structure

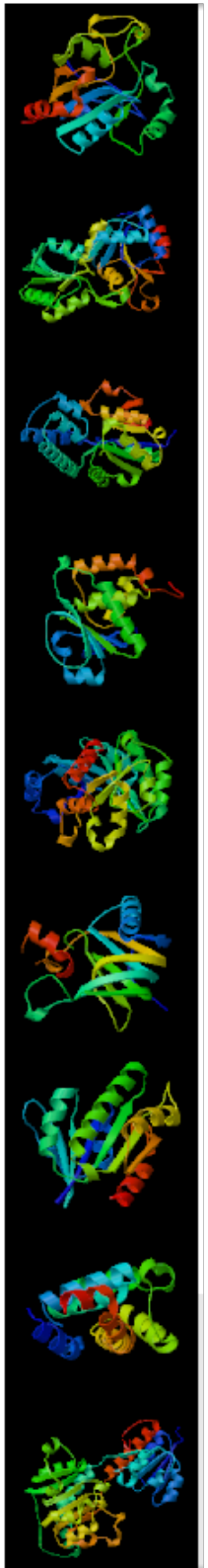
Mounted sequence



Modeled structure

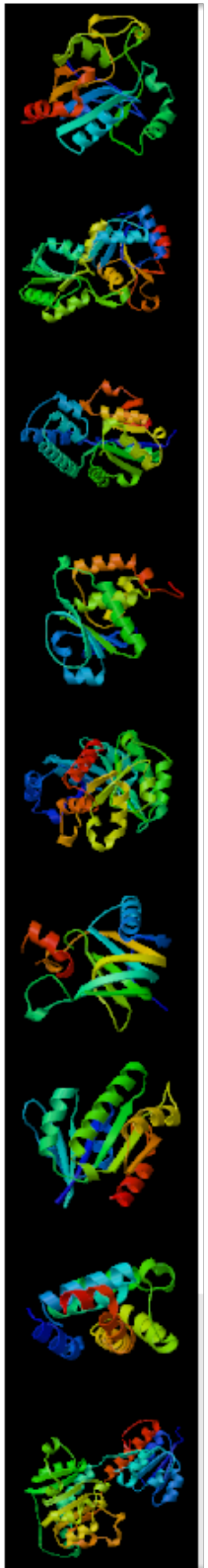


Corrected Model



Evaluating your model

- inaccurate if atomic coordinates are not within 0.5 A RMSD of template control
- Quality checks
  - Bond length
  - Bond Angles
  - Ramachandran
- Biology
  - Does it make sense



Homology Modeling: Practical guide

Approach 1: Manual

- Submit target sequence to BLAST;
  identify potential templates

- For each template:

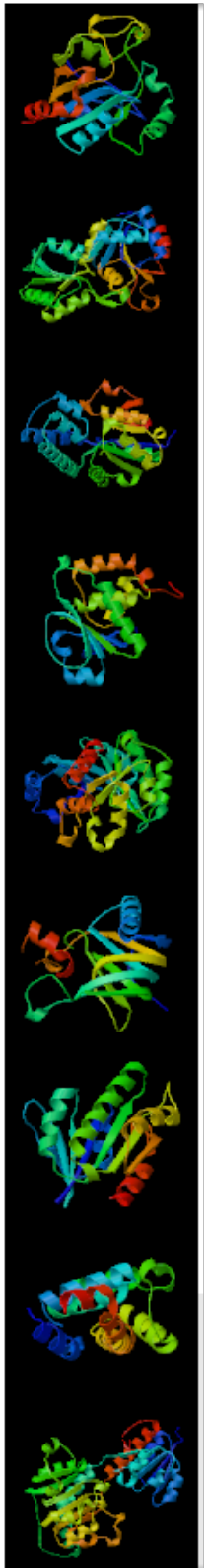
  - Generate alignment between target and template
    (Smith-Waterman + manual correction)

  - Build framework

  - build loop + sidechain

  - assess model (stereochemistry, …)

Arne Elofsson (arne@bioinfo,se)



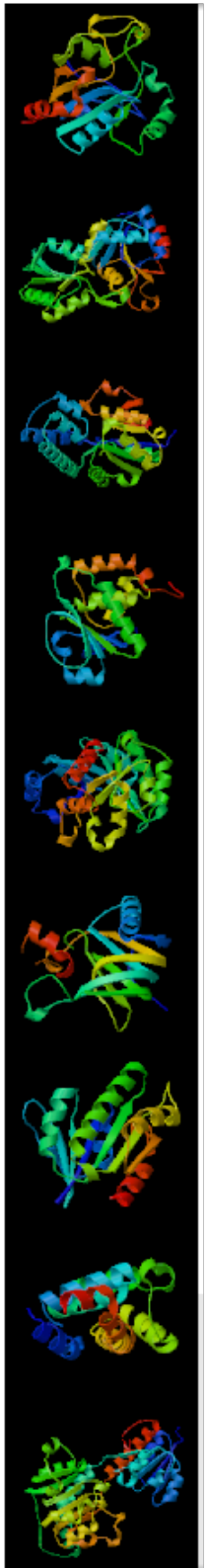
Homology Modeling: Practical guide

Approach 2: Submit target sequence to automatic servers

- *Fully automatic*:
  - 3D-Jigsaw : http://www.bmm.icnet.uk/servers/3djigsaw/
  - EsyPred3D: http://www.fundp.ac.be/urbm/bioinfo/esypred/
  - SwissModel: http://swissmodel.expasy.org//SWISS-MODEL.html
  - Pcons: http://www.cbr.su.se/pcons/

- *Fold recognition*:
  - 3D-PSSM: http://www.sbg.bio.ic.ac.uk/~3dpssm/

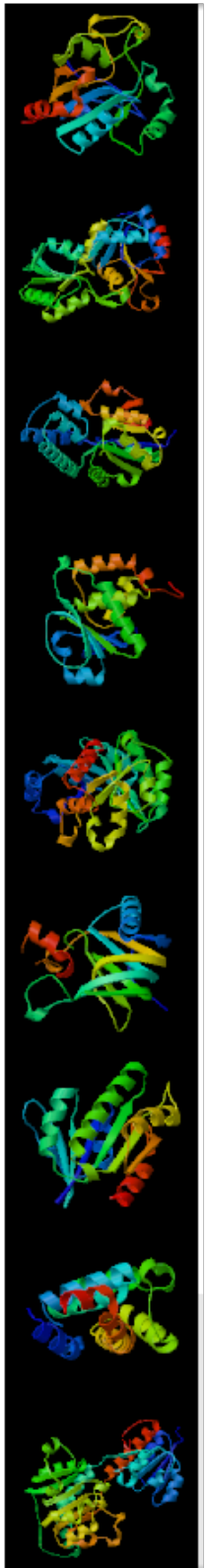
- *Useful sites:*
  - Meta server: http://bioinfo.pl/Meta
  - New Meta server: http://Pcons.net
  - PredictProtein: http://cubic.bioc.columbia.edu/predictprotein/



Homology Modeling: How it works

- *Find template*
- *Align target sequence with template*
- *Generate model:*
  - add loops
  - add sidechains
- *Refine model*

Arne Elofsson (arne@bioinfo,se)



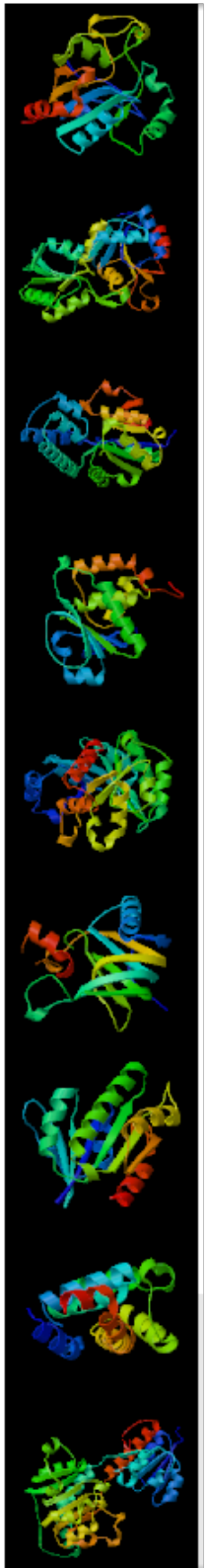
CASP

(Critical Assesment of Structure Predictions)

- the biannual “competition” in protein structure prediction.
- CASP8 next summer

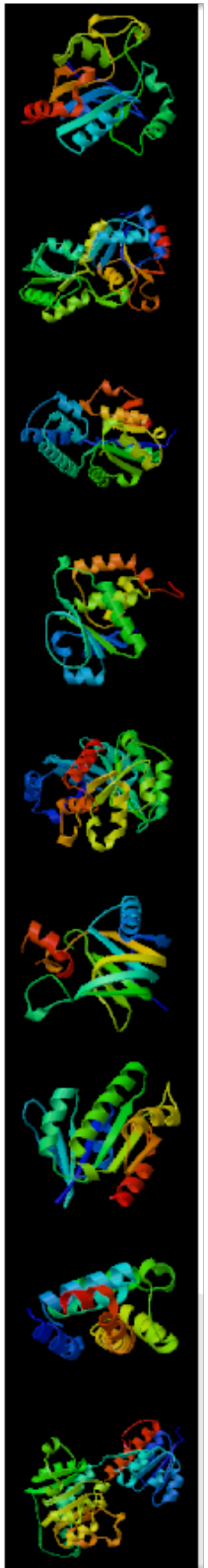
http://predictioncenter.org/

Arne Elofsson (arne@bioinfo,se)



CASP Experiment

- Experimentalists are solicited to provide information about structures expected to be soon solved
- Predictors retrieve the sequence from prediction center (predictioncenter.llnl.gov)
- Deposit predictions throughout the season
- Meeting held to assess results

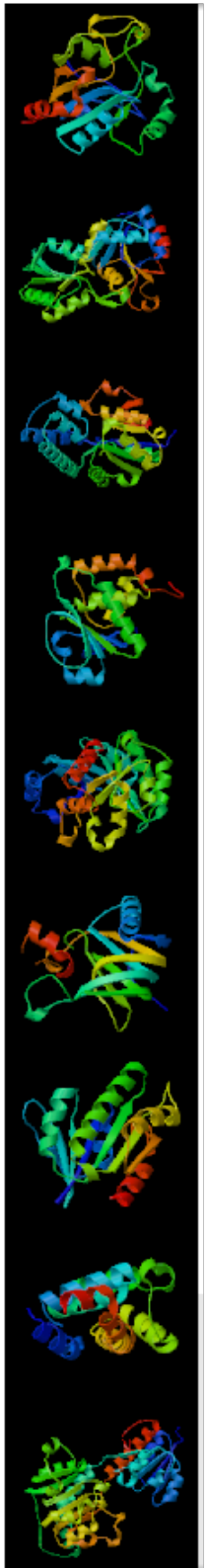


Prediction Categories

- Comparative Modeling – modeling by homology
- Fold Recognition
  - Advanced Sequence Comparison Methods
  - Threading
- New Fold Methods/ “ab initio”
- Categories are separated by distance from any known structure

Conclusions

- When a suitable template structure exists in PDB, using homology modeling on target sequence is best for predicting the structure
- Fold Recognition servers can help find a template when conventional sequence analysis methods fail
- Combining elements from several sources may allow you to construct reasonably accurate models



Summary

- Secondary structure predictions
  - Best methods use machine learning approaches and evolutionary information
  - Close to 80% accuracy
- Homology modeling
  - Needed for the assignments of structure to sequence
  - Good models if %ID is > 50% with automatic methods
- More in the “molecular modelling course”